【临床教学研究】

基于自然语言处理信息抽取技术构建脾胃病科门诊医案语库的探索性研究*

刘瑜[1]　段炳琪[2]　蔺苗苗[3]　姜恩昌[2]　司超增[4]　张喆[1#]

（1 中日友好医院中医脾胃病科 北京 100029；2 北京中医药大学中日友好临床医学院 北京 100029；
3 北京大学第一医院中医、中西医结合科 北京 100034；4 中日友好医院信息部 北京 100029）

摘要：目的　针对中医临床教学中真实世界病历结构杂乱、难以直接转化为教学资源的问题，利用自然语言处理（natural language processing，NLP）技术构建门诊医案结构化语库，旨在为基于案例的中医临床教学和经验传承提供高质量的数字化教具库与素材支撑。**方法**　选取中日友好医院中医脾胃病科主任医师2021年1月1日—2024年12月31日的门诊病历，统一数据口径后使用Python将原始数据进行病历文书清洗和标准化处理，再对同一次就诊下分散保存的主诉、现病史等文书记录进行患者级和就诊级整合，排除没有中医处方且相关内容均为空值的记录，保留可进入病程重构与纳入筛选的可分析就诊记录，研究按照“文本预处理→Prompt构建→调用Qwen3-235B→返回结构化JSON→结果校验→异常重跑或人工复核”的流程提取症状、证型、方药等33个结构化字段，并进行初诊与复诊的病程纵向拆分，最后对结构化内容进行专家级人工验证。**结果**　包含56 949条病历文书片段经就诊级归并形成24 116次门诊就诊记录，覆盖9 709名患者；经字段规范化、自然语言信息抽取、诊断标准化映射、病程拆分和质量控制后，最终形成4 689份有效病案，其中初诊处方字段有值4 672份，复诊处方字段有值2 240份，三诊处方字段有值1 184份。全部4 672份初诊字段有值病案均纳入结构化医案语库建设、病种谱统计及症状-证型-方药关联分析，其中从总体数据集中抽取400份作为质量评价验证子集。结构化医案语库共识别74类中医病名，胃痞1 045例（22.37%）、吐酸780例（16.70%）、胃痛744例（15.92%）居前3位；高频症状包括腹胀、反酸、胃胀、腹部怕冷、嗳气和胃痛；高频中药包括砂仁、桂枝、白芍、陈皮、北柴胡、茯苓、麸炒白术、法半夏和厚朴；关联分析显示，胃痞多与胃胀、腹胀及砂仁、陈皮、厚朴相关，吐酸多与反酸及砂仁、煅瓦楞子相关，泄泻多与腹泻、大便不成形及麸炒白术、茯苓相关。400份验证集由2位中医脾胃科主任医师交叉进行人工复核，整体准确率为87.3%（95%*CI*：83.8%～90.1%），充分满足中医临床教学对医案真实性与严谨性的要求。**结论**　基于NLP技术构建的结构化医案语库可将原本碎片化的中医门诊病历转化为清晰的“理法方药”教学图谱与动态辨证轨迹，能够极大丰富中医临床循证教学与案例教学资源，为医学生临床思维训练、师承教育的数字化发展提供重要支撑。

关键词：中医临床教学；医案语库；自然语言处理；中医传承；案例教学

中图分类号：G642.0　　**doi：**10.3969/j.issn.1003-305X.2026.10.475

Exploratory study on constructing a medical case corpus for TCM spleen and stomach outpatient clinics based on NLP information extraction technology*

LIU Yu[1], DUAN Bingqi[2], LIN Miaomiao[3], JIANG Enchang[2], SI Chaozeng[4], ZHANG Zhe[1#]

(1 Department of TCM for Spleen and Stomach Diseases, China-Japan Friendship Hospital, Beijing 100029, China; 2 Clinical Medical College of Beijing University of Chinese Medicine, China-Japan Friendship Hospital, Beijing 100029, China; 3 Department of Integrated Traditional

刘瑜，女，博士，副主任医师

#通信作者：张喆，男，博士，主任医师，硕士生导师，E-mail：zhangzhebj@163.com

*基金项目：北京中医药大学教育教学改革专项课题（No. XJYZ25025）

Chinese and Western Medicine, Peking University First Hospital, Beijing 100034, China;
4 Information Center, China-Japan Friendship Hospital, Beijing 100029, China)

Abstract: Objective To address the problem that real-world medical records are unstructured and difficult to transform directly into teaching resources in traditional Chinese medicine (TCM) clinical teaching, this study constructed a structured corpus of outpatient medical cases using natural language processing (NLP) technology, aiming to provide a high-quality digital library of teaching aids and material support for case-based TCM clinical teaching and experience inheritance. **Methods** The data originated from the outpatient medical records of a chief physician of the Department of TCM for Spleen and Stomach Diseases at the China-Japan Friendship Hospital from January 1, 2021 to December 31, 2024. After unifying the data caliber, the original records were cleaned and standardized with Python, and the chief complaints, history of present illness, and other documents associated with the same visit were integrated at both the patient and visit levels. Records lacking a TCM prescription or containing only null values were excluded, and only analyzable visits suitable for disease-course reconstruction and screening were retained. A systematic workflow of "text preprocessing→prompt construction→invoking Qwen3-235B→returning structured JSON→result verification→re-running exceptions or manual review" was adopted to extract 33 structured fields, including symptoms, syndrome types, formulas, and medications, followed by longitudinal segmentation of disease courses into initial and follow-up visits and expert-level manual validation of the structured content. **Results** A total of 56 949 medical record fragments were consolidated at the visit level into 24 116 outpatient visit records covering 9 709 patients. After field standardization, natural language information extraction, standardized diagnosis mapping, disease-course segmentation, and quality control, 4 689 valid medical records were finally obtained, of which 4 672 had values in the initial-visit prescription field, 2 240 in the follow-up prescription field, and 1 184 in the third-visit prescription field. All 4 672 records with valid initial-visit fields were included in the construction of the structured medical case corpus, disease-spectrum statistics, and symptom-syndrome-prescription correlation analysis, among which 400 records were selected as a verification subset for quality evaluation. A total of 74 categories of TCM disease names were identified in the corpus, with gastric fullness (1 045 cases, 22.37%), acid regurgitation (780 cases, 16.70%), and stomachache (744 cases, 15.92%) ranking in the top three. High-frequency symptoms included abdominal distension, acid regurgitation, gastric distension, abdominal aversion to cold, belching, and stomachache. High-frequency Chinese materia medica included Amomum villosum, Cinnamomum cassia twig, white peony root, tangerine peel, Bupleurum chinense, Poria, bran-fried Atractylodes macrocephala, Pinellia ternata, and Magnolia officinalis. Correlation analysis showed that gastric fullness was mainly associated with gastric distension, abdominal distension, Amomum villosum, tangerine peel, and Magnolia officinalis; acid regurgitation was mainly associated with acid regurgitation, Amomum villosum, and calcined Arca (Walengzi); and diarrhea was mainly associated with loose stool, bran-fried Atractylodes macrocephala, and Poria. The 400-record verification set was cross-checked manually by two chief physicians of TCM spleen and stomach diseases, with an overall accuracy of 87.3% (95%*CI*: 83.8%~90.1%), fully meeting the requirements of TCM clinical teaching for the authenticity and rigor of medical cases. **Conclusion** The structured medical case corpus constructed with NLP technology can transform fragmented TCM outpatient records into clear teaching maps of "principle, method, formula, and medication" and dynamic syndrome-differentiation trajectories, greatly enriching the resources for evidence-based and case-based teaching of TCM clinical practice, and providing important support for the clinical thinking training of medical students and the digital development of master-apprentice education.

Keywords: traditional Chinese medicine clinical teaching; medical case corpus; natural language processing; traditional Chinese medicine inheritance; case-based learning teaching

中医药临床人才的培养高度依赖真实医案的研读与名老中医经验的师承。真实的门诊病历不仅记录了诊疗经过,更承载了专家临证时的辨证思维、症状演变规律和处方加减逻辑,是极其宝贵的教学资源。然而,医院信息系统沉淀下来的原始资料往往以业务视角组织,常表现为多文件、多表格、多主题并存,且主诉、现病史、体格检查、辨证处方等核心教学内容以自然语言形式分散存储。这种异质性强、结构杂乱的数据状态导致医学生在自主学习时难以从海量病历中高效检索和提炼出清晰的"理、法、方、药"逻辑链条,使得传统的教学模式面临严重的效率瓶颈[1-3]。

本研究从中日友好医院中医脾胃病科门诊病历出发,经Python预处理、病历自然语言特征提取、结构化查询语言(structured query language,SQL)重构,建立一条能够追溯结果来源的完整技术链路,其总体思路与既往电子病历结构化研究的技术路线一致,但本研究更强调针对中医门诊教学场景,实现从"原始零散文本"到"结构化教学语库"的全链路闭环转换[2-4]。本研究期望将海量中医临证经验转化为可检索、可追踪的标准化教学资产,为中医临床循证教学(evidence-based teaching,EBT)和基于案例的教学(case-based learning,CBL)提供高质量的数据基座。

1　资料与方法

1.1　研究资料数据口径界定

选取中日友好医院中医脾胃病科主任医师2021年1月1日—2024年12月31日的门诊病历。为避免不同统计单位混用,本研究对数据口径作如下界定:"文书"指医院信息系统中分散保存的主诉、现病史、诊断、处方等原始文书片段;"就诊记录"指同一患者在某一门诊日期形成的一次诊疗记录;"患者"指经脱敏后由唯一标识符关联的个体;"病案"指经纳入标准筛选、完成结构化抽取和质量控制后进入分析的患者病程单元;"病程"指同一患者按就诊时间连续追踪的诊疗序列,当前后2次就诊间隔超过6个月时,判定为新的病程起点。

1.2　多源异构原始数据整合与主题表建设

为把碎片化病历整理成能稳定入库和提取的标准输入,首先使用Python对原始门诊病历文书进行清洗和标准化处理,包括统一列名、时间格式、空值表示、换行符和编码形式,并保留原始来源信息。再对同一次就诊下分散保存的主诉、现病史等文书记录进行患者级和就诊级整合。

完成清洗后,将病历数据导入PostgreSQL主题表,在数据库层保留原始颗粒度,避免入库阶段过度汇总。研究中以患者标识和就诊记录为基础建立关联索引,为后续病程拆分、特征拼接和结果追溯建立统一基础。随后排除没有中医处方且相关内容均为空值的记录,保留可进入病程重构与纳入筛选的可分析就诊记录。经预处理(提取至少6个月内有2次及以上就诊,主诉、诊断等核心说明没有丢失,且有中医处方的患者)、特征提取、结构化重构及质量控制后,最终得到的有效病案构建为初诊特征统计宽表,以此作为后续统计分析的核心样本。具体见图1。

1.3　实施病历自然语言特征提取

病历文书中的中医关键信息大多表现为非结构化自然语言,尤其是主诉、现病史、体格检查和诊断中涉及的大量症状、睡眠饮食情况、二便信息、舌脉特征及既往病史等内容,难以直接用于统计分析[4-6]。对此,本研究采用云端API调用Qwen3-235B-A22B模型完成病历文本结构化抽取。为提高输出稳定性,模型调用采用固定Prompt模板和低随机性参数设置:temperature=0.1,top_k=3,其余参数保持云端API默认配置。所有输入文本在调用前均完成脱敏处理,不包含患者姓名、身份证号、联系电话、门诊号等直接识别信息。

Prompt由任务说明、字段定义、缺失值规则、JSON输出格式和患者文本5个部分组成。模型被要求仅根据输入病历内容抽取信息,不得自行补充或推断未出现内容;未提及字段统一返回空字符串,原文明确无法判断时返回"-";输出必须为合法JSON,不得输出解释性文字。提取字段包括:入院诊断、主要诊断、次要诊断1、次要诊断2、次要诊断3、证候、主诉中的症状1、主诉中的症状时间1、主诉中的症状2、主诉中的症状时间2、诱因、症状、性质、程度、伴随症状、持续或间断性、加重或缓解、精神状态、食欲、饮食偏好、冷热喜恶、大便是否正常、大便次数、大便性状、便意、睡眠情况、舌色、舌形、舌边齿痕、舌苔部位、舌苔厚薄、舌苔颜色和脉象等33个结构化字段。具体见图2。需要说明的是,原始文书中的入院诊断为单次就诊记录中的所有诊断,主要诊断指单次就诊记录中的第一诊断,次要诊断1为第二诊断,依次为次要诊断2、3,证候为原始文书中的中医证型,主诉中的症状1为单次就诊中患者最不适或最亟待解决的症状,主诉中的症状2为主诉中的次要症状或伴随症状,后续字段均按

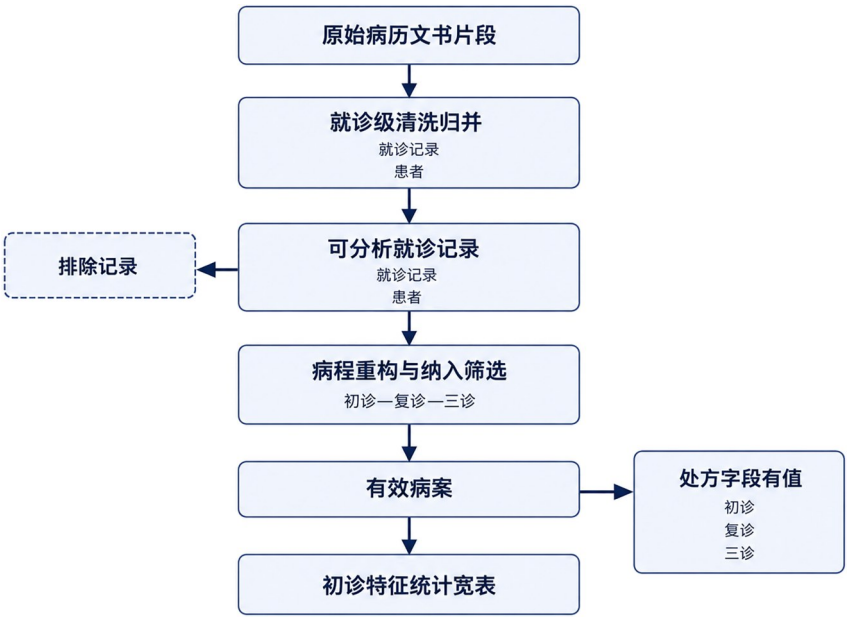


图 1　研究数据流向图

任务描述与模型参数

你是一个医学病历结构化信息提取助手。请根据输入的中医脾胃病门诊病历文本，抽取单次就诊的结构化字段。

模型：qwen3-235b；部署方式：云端 API；
temperature=0.1；top_k=3；max_tokens=平台默认值。

抽取要求

1. 仅依据输入原文抽取，不得自行编造、补充或医学推断。
2. 原文未提及的字段返回空字符串；原文明确无法判断时返回“-”。
3. 输出必须为合法JSON，不得输出解释、Markdown或多余文字。
4. 字段名必须固定，不能改名、合并或新增字段。
5. 可在evidence_text中保留关键证据片段，便于人工复核。
6. 若输入包含初诊、复诊、三诊，应分别输出；解析为宽表时按就诊阶段加前缀，如“初诊症状”“复诊症状”“三诊症状”。

输出 JSON 模板（节选）

```
{
  "visit_stage": "",
  "extracted_fields": {
    "入院诊断": "",
    "主要诊断": "",
    "次要诊断1": "",
    …
    "症状": "",
    "伴随症状": "",
    …
    "舌苔颜色": "",
    "脉象": ""
  },
  "evidence_text": {}
}
```

患者文本

```
{patient_text}
```

注：“…”表示省略的字段；完整模板含33个抽取字段。

图 2　Prompt 示例

照原始文书中的现病史、既往史描述进行医学级拆解。33个结构化字段按临床教学用途分为诊断类、主诉症状类、病程与诱因类、症状性质与伴随症状类、饮食二便睡眠类、舌脉类6组;其中诊断类字段用于病名谱和案例筛选,症状及舌脉字段用于辨证训练,方药及剂量字段用于处方讨论和复诊加减分析。

模型返回结果后,程序依次进行JSON合法性校验、字段名一致性校验、字段完整性校验和异常值识别。若返回内容无法解析为合法JSON,或字段名与模板不一致,则记录异常并按同一Prompt重新处理;若重新处理后仍不合格,则转入人工复核列表。对原文未提及且模型返回为空的字段按缺失值处理;对少量可由固定文本规则识别的舌脉字段,在保留模型结果的基础上进行规则回填并记录来源。该做法综合了中文电子病历命名实体识别、临床概念抽取以及大语言模型结构化抽取的常见范式[7-15]。具体见图3、表1。

输入病历

反流性食管炎间断发作7年余，夜间严重，时有上腹部疼痛，全身怕冷，时有大便干，每日一次，咽部疼痛，舌质紫暗，有瘀点，舌苔白腻，脉沉滑。既往咽部手术，既往甲状腺功能减低，既往子宫术后。

模型与参数

模型：qwen3-235b；就诊阶段：初诊；
temperature=0.1；top_k=3；max_tokens=平台默认值。

结构化输出（节选）

```
{
  "case_id": "case_tusuan_001",
  "visit_stage": "初诊",
  …
  "extracted_fields": {
    "入院诊断": "反流性食管炎；便秘",
    "主要诊断": "吐酸",
    "次要诊断1": "便秘",
    …
    "程度": "严重",
    "伴随症状": "上腹部疼痛、全身怕冷、大便干、咽部疼痛",
    "持续或间断性": "间断",
    …
    "冷热喜恶": "怕冷",
    "大便是否正常": "不正常",
    "大便次数": "1次",
    "大便性状": "干",
    …
    "舌色": "紫暗",
    "舌苔厚薄": "厚腻",
    "舌苔颜色": "白",
    "脉象": "沉滑"
  },
  "evidence_text": {
    …
    "舌脉": "舌质紫暗，有瘀点，舌苔白腻，脉沉滑"
  },
  …
}
```

注：仅展示33个抽取字段中的部分字段；“…”表示省略内容。

图3　JSON输出示例(吐酸复杂现病史,33字段)

经上述处理后,自然语言病历被转换为可入库的结构化特征表,从而实现对症状、时间、诱因、饮食睡眠、二便、舌脉及既往病史等多维信息的统一抽取;这一“模型抽取+规则校验+异常回填”的混合流程,与既往临床信息抽取系统的质量控制思路一致[16-18]。

表1　33字段映射与本案例输出

字段	宽表列名	本案例输出
入院诊断	初诊入院诊断	反流性食管炎;便秘
主要诊断	初诊主要诊断	吐酸
次要诊断1	初诊次要诊断1	便秘
次要诊断2	初诊次要诊断2	
次要诊断3	初诊次要诊断3	
症候	症候	
主诉中的症状1	初诊主诉中的症状1	–
主诉中的症状时间1	初诊主诉中的症状时间1	–
主诉中的症状2	初诊主诉中的症状2	–
主诉中的症状时间2	初诊主诉中的症状时间2	–
诱因	初诊诱因	–
症状	初诊症状	反流性食管炎、上腹部疼痛、全身怕冷、大便干、咽部疼痛、舌质紫暗、有瘀点
性质	初诊性质	–
程度	初诊程度	严重
伴随症状	初诊伴随症状	上腹部疼痛、全身怕冷、大便干、咽部疼痛
持续或间断性	初诊持续或间断性	间断
加重或缓解	初诊加重或缓解	–
精神状态	初诊精神状态	–
食欲	初诊食欲	–
饮食偏好	初诊饮食偏好	–
冷热喜恶	初诊冷热喜恶	怕冷
大便是否正常	初诊大便是否正常	不正常
大便次数	初诊大便次数	1次
大便性状	初诊大便性状	干
便意	初诊便意	–
睡眠情况	初诊睡眠情况	–
舌色	初诊舌色	紫暗
舌形	初诊舌形	–
舌边齿痕	初诊舌边齿痕	–
舌苔部位	初诊舌苔部位	–
舌苔厚薄	初诊舌苔厚薄	厚腻
舌苔颜色	初诊舌苔颜色	白
脉象	初诊脉象	沉滑

1.4　诊断标准化与患者纵向病程重构

将电子病历的诊断字段拆分为西医诊断和中医诊断:西医诊断拆分为主诊断和次要诊断,中医诊断拆分为中医病名与证候,并连接诊断归属表完成标准化映射。通过这一过程,原始书写方式不一的诊断文本被统一到可比较、可统计的标准字段中,并可进一步开展病名频次统计与病种谱分析。这也是非结构化病历向研究型数据库转化的关键步骤[14]。

随后按同一患者关联记录内的就诊时间排序,若与前次就诊间隔超过6个月,则判定为新的病程起点并生成病程标识;同时利用就诊序号标记病程内的初诊、复诊和三诊顺序。这样便把原本离散的多次就诊记录重建为可纵向追踪的病程数据,为分析症状演变、处方加减和剂量变化提供了数据基础;该类纵向重构也符合电子病历研究中从就诊级病程转向患者级病程分析的常见框架[14-15]。

1.5　研究数据集生成与指标整理

在完成特征提取和病程重构后,进一步生成用于统计分析和机器学习建模的研究数据集,并据此构建门诊医案语库数据库。数据库设计包括核心表与关系表2个部分。核心表包括:病例主表,用于存储患者标识符(identifier,ID)、就诊时间等基本索引信息;症状表,用于存储症状名称、部位、程度等结构化症状信息;证型表,用于存储证型名称及其辨证依据;方药表,用于存储方剂组成及对应剂量。关系表包括:症状-证型关联表,用于记录症状与证型之间的对应关系及置信度权重;证型-方药推荐表,用于记录证型与方药之间的关联频次统计。与此同时,对历次处方进行研究性过滤,排除大量非研究药物、辅药、西药和治疗操作项目,并去除剂量为空或0的记录;同时整理药物是否使用、药物剂量、病程特征及其他结构化变量,使其能够直接进入关联规则分析、列联检验、分类模型训练和特征重要性评估。至此,自然语言特征、标准化诊断和处方信息被汇总为面向分析的结构化宽表,并进一步组织为可查询、可关联、可扩展的医案语库数据库。具体见图4。

1.6　验证集构建与人工抽检

为评价基于结构化语料库的自然语言处理(natural language processing,NLP)抽取结果是否达到教学和临床可用水平,本研究并未另行构建结构不同的数据集,而是在有效病案构成的总体研究数据集中抽取质量评价子集。该验证集仅用于人工复核和准确率估计;抽检完成后,其结构化结果仍作为总体语库的一部分。除验证集病案外,其余有效病案同样纳入结构化医案语库建设、病种谱与病名统计、症状—证型—方药关联分析、病程纵向分析及教学案例资源构建。验证重点评估3类内容:基础结构字段,包括患者ID、就诊时间、病程序号

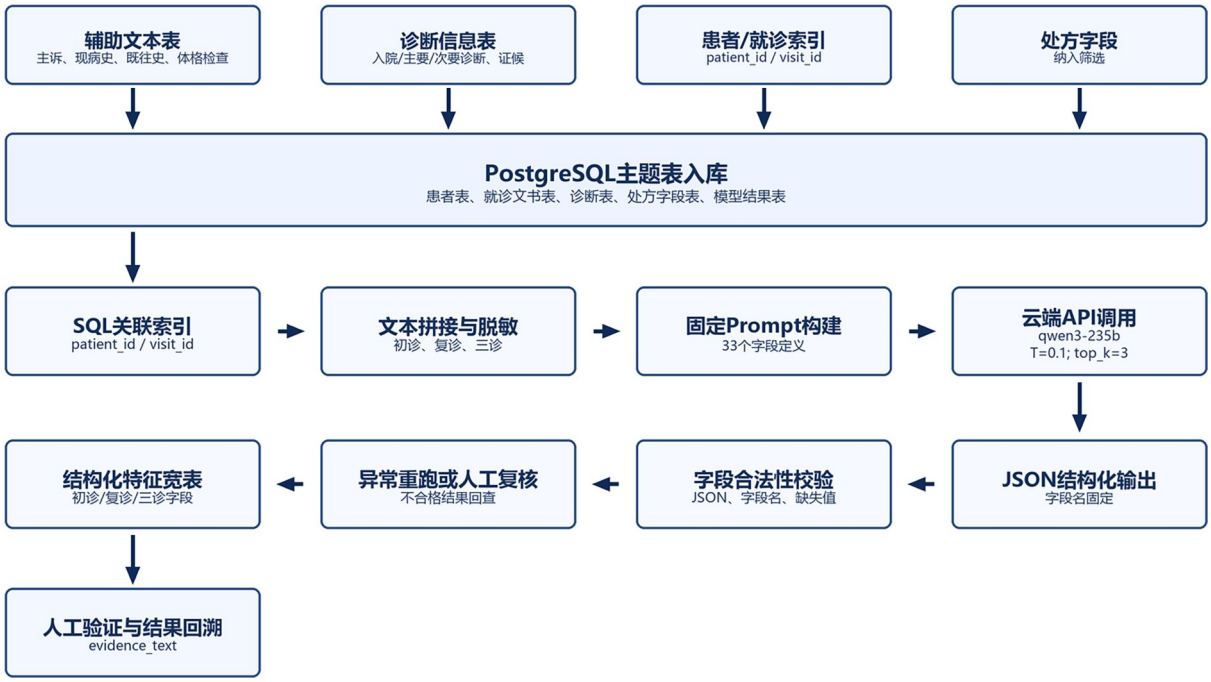


图 4　病历自然语言特征提取与 PostgreSQL 重构流程

等;临床语义字段,包括症状名称、症状持续时间、诱因、性质、伴随症状、饮食睡眠、二便、舌脉、既往史及证型;方药字段,包括方剂组成、单味药剂量及症状—证型、证型—方药关联结果。验证集构建采用分层随机抽样,而非简单随机抽样。分层维度包括:就诊阶段(初诊、复诊、三诊)、文本复杂度(短文本、中等文本、长文本)、抽取任务类型(症状类、时间类、舌脉类、既往史类、证型类、方药类),以及结果状态(完整抽取、部分缺失、模型低置信或规则回填病例),以避免验证样本仅覆盖“容易病例”,保证样本代表性。样本量按准确率估计公式 $n=Z^2\times p(1-p)/d^2$ 计算,其中预期准确率 p 取 0.85,允许误差 d 取 0.04,检验水准 α 取 0.05,对应 Z 值为 1.96 来估算最小样本量,考虑分层抽样、复核损耗及专家意见不一致等情况,最终计算出验证集样本量。由 2 位中医脾胃病科主任医师采用交叉独立盲法进行人工审核,每位专家仅查看原始病历文书,不查看模型结果来源说明,根据统一标注表对每个字段判定为“正确”“部分正确”“错误”或“无法判断”。若 2 位专家意见不一致,则通过进一步讨论形成一致意见,并以该一致意见作为人工金标准。

字段判定标准按类型分别设定:症状类字段以抽取结果与原文语义一致为正确,同义术语如“胃脘痛”与“胃痛”按预设词表归一后可判定为正确;时间类字段要求时间长度或时间表述与原文一致,“3 天”与“三日”等可归一;舌脉类字段要求舌色、舌形、苔色、苔厚薄、脉象与原文一致,若原文未提及而系统返回空值亦判定为正确;证型类字段要求证型名称符合原文辨证结论或能够依据辨证依据唯一推断;方药类字段要求药名一致且剂量一致,药名一致但剂量偏差超出预设阈值者判为错误;关联类字段不逐条验证全部关联边,而是抽取高频症状-证型关联及证型-方药关联结果,由专家进行合理性审核。统计学评价包括 2 个部分:首先评价 2 位专家之间的一致性,分类变量采用 Cohen’s kappa,有序变量采用 weighted kappa,并按<0.40 为一致性较差、0.40 ~ 0.75 为中等一致、>0.75 为一致性较好进行解释;其次以专家最终一致意见为金标准,计算 Accuracy 及其 95% 可信区间,并结合 Kappa 值评价抽取结果的稳定性和专家一致性。

2　结果

2.1　自然语言病历结构化提取结果

原始资料共包含 56 949 条病历文书片段,经就诊级归并形成 24 116 次门诊就诊记录,覆盖 9 709 名患者;排除无中医处方或关键字段均为空值的记录 3 679 次后,保留 20 437 次可分析就诊记录和 7 695 名患者;再经字段规范化、自然语言信息抽取、诊断标准化映射、病程拆分和质量控制后,最终形成 4 689 份有效病案。由于部分病案初诊主要诊断缺失或不满足标准化统计条件,初诊主要诊断分布和初诊结构化字段统计以 4 672 份病案为分母。处方字段有值数仅作为纳入筛选与病案完整性说明,其中初诊

处方字段有值4 672份，复诊处方字段有值2 240份，三诊处方字段有值1 184份。具体见图5、图6。验证集来源于上述4 672份有效病案用于质量评价的抽样子集，而非另行构建的结构不同数据集，根据上述公式估算得出最小样本量为306份，考虑分层抽样、复核损耗及专家意见不一致等情况，最终按分层随机抽样抽取400份有效病案作为验证集；其余约4 272份病案与验证集共享同一结构化字段体系，继续用于语库建设、病名统计、症状—证型—方药关联分析、病程规律挖掘和教学案例资源构建。具体见表2、表3。

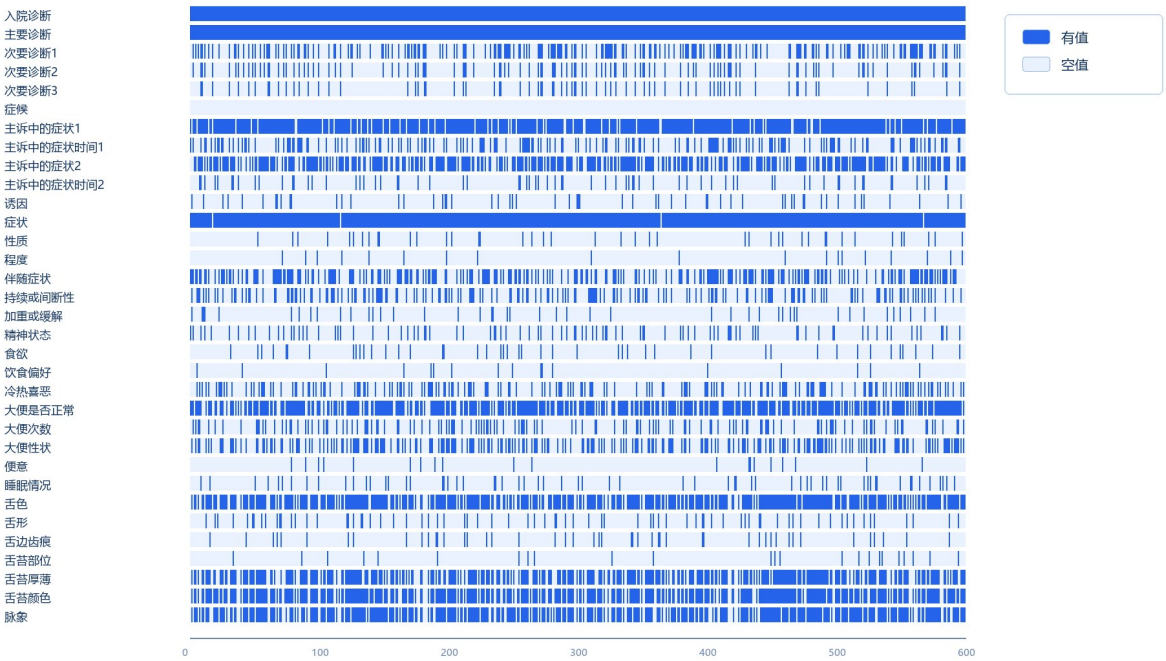


图5　33个提取字段缺失二值矩阵

注：在4672份初诊字段有值病案中随机抽取600份实际病案，每行代表提取字段，每列代表实际病案抽样，图例仅区分有值/空值。

2.2　多源数据重构后的研究数据库特征

基于病名计数结果，在可完成初诊主要诊断标准化统计的4 672份病历中，共归纳74类中医病名。病种分布呈现较明显的头部聚集特征，其中胃痞1 045例(22.37%)、吐酸780例(16.70%)、胃痛744例(15.92%)、泄泻415例(8.88%)和便秘389例(8.33%)位居前5位，合计占72.20%；前10位病名合计占82.75%，提示该结构化语料库能够较好覆盖脾胃病科门诊的主要病种谱。与此同时，腹痛、失眠、嗳气、口臭和咳嗽等病名仍具有一定比例，而脱发、鼓胀、舌炎、身痒、过敏性紫癜等病名各仅1例，说明该语料库在保留门诊高频脾胃病种的同时，也保留了少量长尾病种信息，为后续开展不同病种层面的症状—证型—方药关联分析提供了数据基础。具体见图7。在可完成结构化统计的4 672份病案中，高频症状以腹胀801例(17.14%)、反酸725例(15.52%)、胃胀651例(13.93%)、腹部怕冷463例(9.91%)、嗳气455例(9.74%)、胃痛437例(9.35%)、口干424例(9.08%)、打嗝424例(9.08%)、便秘419例(8.97%)、大便不成形411例(8.80%)较为突出。证型及舌脉相关要素中，苔白2 111例(45.18%)、苔黄1 636例(35.02%)、脉沉滑1 121例(23.99%)、脉沉细990例(21.19%)、脉弦滑674例(14.43%)较为常见，提示痰湿、郁热、气机失调为高频证型。方药层面，高频药物包括砂仁1 986例(42.51%)、桂枝1 711例(36.62%)、白芍1 539例(32.94%)、陈皮1 534例(32.83%)、北柴胡1 526例(32.66%)、茯苓1 441例(30.84%)、麸炒白术1 424例(30.48%)、法半夏1 387例(29.69%)、厚朴1 312例(28.08%)、肉桂1 185例(25.36%)、瓜蒌1 168例(25.00%)、丁香1 114例(23.84%)、香附1 093例(23.39%)、黄芩1 053例(22.54%)和白术1 046例(22.39%)。这些药物共同覆盖理气和胃、温通降逆、健脾化湿、清热化痰和调和肝脾等治疗方向，为“理法方药”教学图谱提供了可量化的药物节点。

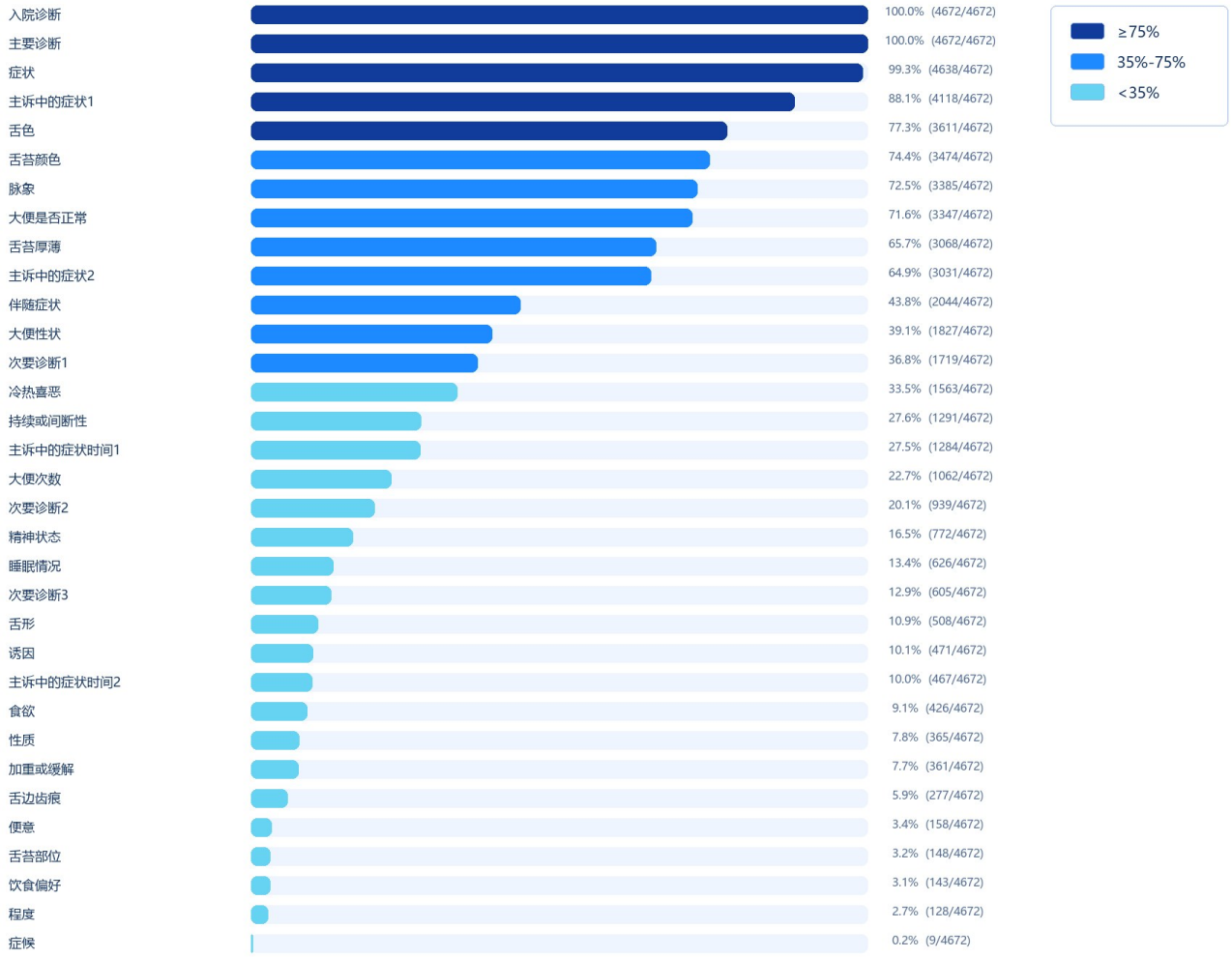


图 6　33 个提取字段非空率排序

注：按照 4672 份初诊字段有值病案特征统计样本计算。

表 2　验证集构建与抽样分层情况表

（*n*=400，随机种子：20260413）

分类	例数
就诊阶段	
初诊	209
复诊	90
三诊	101
文本复杂度	
短文本	135
中等文本	133
长文本	132
任务类型	
症状类	399
时间类	365
舌脉类	344
既往史类	139
证型类	25
方药类	400

表 3　就诊阶段与文本复杂度交叉分布

就诊阶段	短文本	中等文本	长文本
初诊	125	77	7
复诊	6	42	42
三诊	4	14	83

主要病种的关联示例显示，胃痞与胃胀 349 例、腹胀 317 例、砂仁 500 例、陈皮 479 例、厚朴 450 例、法半夏 408 例关联较高；吐酸与反酸 362 例、嗳气 174 例、砂仁 362 例、煅瓦楞子 234 例关联较高；胃痛与胃痛症状 251 例、反酸 128 例、桂枝 340 例关联较高；泄泻与腹泻 127 例、大便不成形 107 例、麸炒白术 368 例、茯苓 302 例关联较高；便秘与便秘症状 190 例、白芍 297 例关联较高。上述结果说明，该语库不仅能够描述病种频次，还可以进一步呈现“病名—症状—方药”的多条关联关系，为后续 CBL 病例筛选和师承教学中的辨证路径展示提供数据基础。

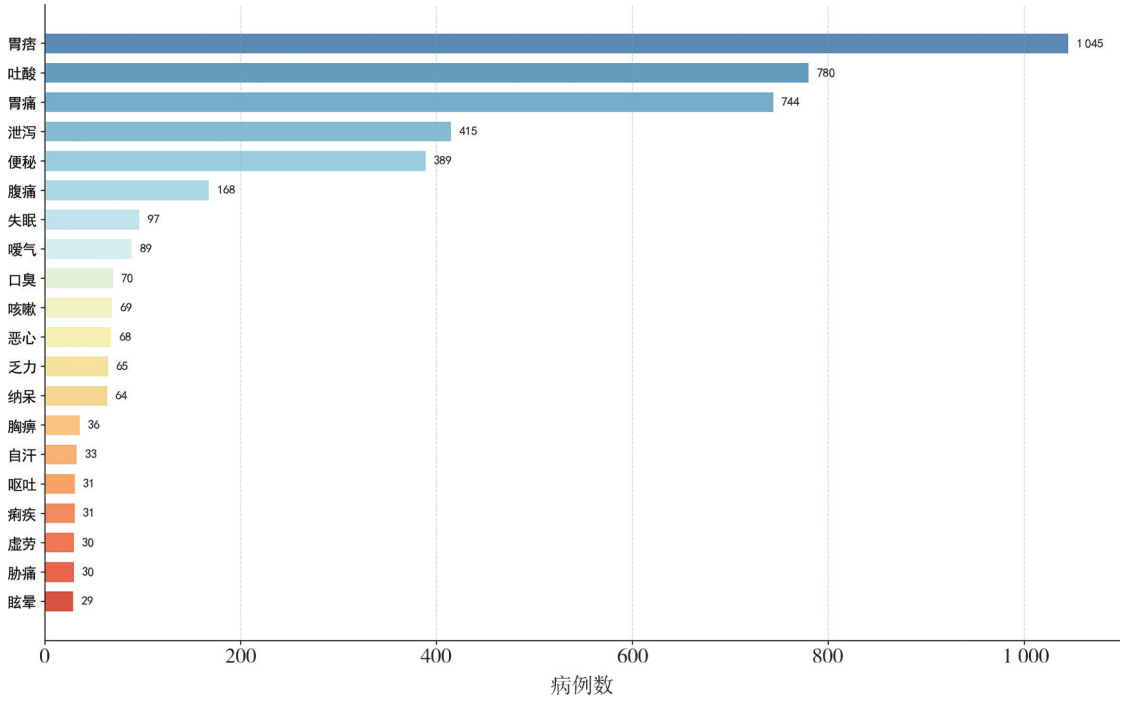


图 7　主诊断分布图(前 20 位)

2.3　验证集人工抽检结果

人工抽检结果显示，从 4 672 份有效病案中按分层随机抽样纳入的 400 份质量评价子集，经 2 位中医脾胃病科主任医师独立审核后，多数核心字段能够与原始病历文书保持一致。该验证集与总体研究数据集为子集与总体的关系，仅承担抽取质量评价功能，并不改变与其余约 4 272 份病案共享的字段结构和语库用途。从字段类型看，症状名称、舌脉特征、既往史及方药组成等信息的一致性较高；病程、症状持续时间及跨句描述的伴随症状在少数病例中存在偏差。以 2 位专家最终一致意见作为人工金标准计算，整体准确率为 87.3%(95%*CI*：83.8%～90.1%)，提示该结构化提取方法已达到较好的临床和教学可用水平。专家间一致性评价显示，核心字段审核结果一致性良好。误差主要来源于原文表述含混、跨句省略、时间描述不完整，以及个别临床术语归一困难等情况。需要说明的是，验证集中证型字段的一致性结果主要用于发现抽取流程中的潜在错误和提示人工复核重点，而证型类样本量主要受原始病历中证型类字段偏少、空值偏多的影响，后续研究需针对证型阳性样本量增加分层抽取比例，并结合多医师、多专科资料进一步评价代表性。具体见表 4、表 5。

表 4　人工抽检结果表

字段类别	样本数	准确率(%)	95%*CI*	Kappa 值
总体字段抽检	400	87.3	83.8～90.1	0.82
基础结构字段	400	96.0	94.1～97.9	0.90
症状类字段	399	87.0	83.7～90.3	0.82
时间类字段	365	82.5	78.6～86.4	0.76
舌脉类字段	344	85.5	81.8～89.2	0.79
既往史类字段	139	82.0	75.6～88.4	0.74
证型类字段	25	82.0	66.9～97.1	0.73
方药类字段	400	87.0	83.7～90.3	0.82

2.4　结构化数据库构建结果

结果显示，病例主表承担患者 ID 和就诊时间等索引管理功能；症状表沉淀症状名称、部位和程度等细粒度表现；证型表记录证型名称及其辨证依据，方药表记录方剂组成和剂量信息；两类关系表则进一步描述症状与证型之间的置信度联系及证

表5　错误类型分析表

错误类型	计数	错误率(%)	备注
漏抽	13	3.2	多见于伴随症状或既往史信息遗漏
错抽	8	2.0	主要为否定语境或非症状表达被误识别
同义词归一失败	7	1.8	集中于症状术语与中医表述归一不充分
跨句信息遗漏	10	2.5	常见于现病史跨句关联和转折表达
时间描述错误	8	2.0	主要涉及病程时间粒度或时间单位解析
剂量解析错误	5	1.2	主要为方药剂量拆分和规格识别偏差

型与方药之间的关联频次。这种“核心表+关系表”的设计,使门诊医案语库不仅能够支持单条病历检索,还能够支撑“症状-证型-方药”之间的多条关联分析,为后续知识挖掘、规则发现和临床决策支持的EBT和CBL教学奠定数据库基础。

2.5　方法的可迁移性

本文方法强调的是“真实世界原始数据整合—病历自然语言特征提取—病程级结构化重构—分析结果输出”的通用过程,而非局限于某一单病种或单一医师文书风格。因此,从技术路径上看,该方法不仅适用于当前数据集,也可在诊断映射规则、验证方案和分析指标体系适度调整后扩展至其他中医专科场景。

3　讨论

3.1　NLP抽取是本研究结果产生的关键环节

本研究中Qwen3-235B-A22B模型承担的是字段级病历特征提取任务,其输出需经过JSON校验、异常复核和规则回填;随后还要经过诊断标准化、病程拆分、症状标注和处方过滤等一系列后处理步骤,才能进入最终统计分析。尤其是在本研究中,模型不仅提取主要诊断、舌脉等传统字段,还对主诉症状及时间、病程、诱因、饮食睡眠、二便、既往史等多维临床信息进行统一抽取,进一步通过2位中医脾胃病科主任医师人工抽检验证,核心字段整体准确率为87.3%(95%*CI*:83.8%~90.1%),从而为结构化结果的可信度提供临床依据。因此,自然语言处理并非陪衬性技术,而是连接原始业务数据与科研结论的关键桥梁;这一定位与近年来大语言模型用于电子病历结构化抽取和临床数据整理的研究认识相一致[16-20]。

3.2　从结构化提取走向临床知识组织

本研究的价值不只在于生成一张结构化结果表,更在于把症状、舌脉、诊断、处方和剂量等原本分散在自然语言文本和业务表中的信息重建为可纵向追踪、可关联分析的研究数据结构。病名计数结果进一步说明,该结构化语料库在病种层面具备较好的临床覆盖度:一方面,胃痞、吐酸、胃痛、泄泻、便秘等脾胃病科高频病名占据主体,符合专科门诊实际;另一方面,仍保留少量长尾病名,为后续探索“症状—证型—方药”关系在不同病种中的迁移性提供了基础。在此基础上,以病例主表、症状表、证型表、方药表及其关系表为核心的数据库设计进一步将自然语言处理结果转化为可查询、可统计、可扩展的医案语库。由此,结构化结果可以继续进入热图、网络图、列联分析图、ROC曲线和特征重要性分析等流程,使中医电子病历从原始资料转化为可分析、可复用的研究数据资源;这与中医医案知识组织及大语言模型结合知识图谱的研究方向相呼应[1,19]。

3.3　结构化医案语库对中医临床教学体系的赋能价值

本研究构建的结构化医案语库,为破解中医临床教学中“经验隐性化”和“理论脱离实际”的问题提供了切实可行的数字化路径,为中医教学赋能。一方面,构建全景式“理法方药”知识图谱,丰富CBL教学素材。传统中医教学多依赖教材中的典型医案,而真实世界的患者症状往往错综复杂,本研究通过NLP抽取,将非结构化文本重构为包含病例主表、症状表、证型表、方药表及其关系表的数据库,构建“症状—证型—方药”的多条关联网络,直观揭示脾胃病科高频病种(如胃痞、吐酸等)在真实世界中的辨证规律,帮助学生弥合教材理论与复杂临床实际之间的认知差距。另一方面,重构纵向病程轨迹,实现“动态辨证与处方加减”的显性化教学。中医临证教学的最大难点在于教授学生如何根据患者复诊时的症状进退进行“效不更方”或“随证加减”。本研究在数据处理阶段特意引入了病程拆分机制,对初诊、复诊、三诊进行时间轴序列化排序,将原本离散的就诊记录转化为可纵向追踪的治疗轨迹,医学生可通过检索语库,清晰对比同一患者在不同诊疗阶段的症状演变与药物剂量调整细节,从而深入体会专家“观其脉证,知犯何逆,随证治

之”的动态辨证精髓，极大提升了中医临床思维训练的客观性与直观性。

从教学转化角度看，频次统计结果可进一步转化为案例筛选规则和课堂任务。以“吐酸病”教学为例，可优先筛选病名为吐酸，主诉含反酸或烧心，具有初诊与复诊连续记录，且舌脉和处方信息较完整的病例；学生课前完成症状清单和病机假设，课堂中围绕“反酸、嗳气、胃胀、口苦等症状如何指向寒热错杂、肝胃不和或脾胃虚弱”等问题开展辨证讨论，并比较砂仁、煅瓦楞子、黄连、吴茱萸、瓜蒌等高频药物在不同证候组合中的配伍意义；课后则依据复诊记录评价症状转归、处方加减是否与辨证变化一致。该流程可形成“病例检索—症状归纳—证型判断—方药分析—复诊追踪—形成性评价”的CBL教学闭环，也可延伸至门诊见习和师承教学中的跟诊前预习、跟诊后复盘及病例汇报[17-21]。

3.4 局限性

本研究仍存在一定局限：一是部分症状标签和结局变量的判定仍结合关键词和文本规则，后续仍需进一步提升语义判别精度；二是当前结果主要基于现有研究数据集完成，尚需在更多专科场景和更多任务类型中进行系统验证；三是当前研究重点在于完成从自然语言病历到结构化语库的底层数据基座构建，距离将其直接开发为面向中医学子的智能化教学辅助平台（如智能互动问答、虚拟模拟诊疗系统）仍有拓展空间，这也是本课题组下一步在教育技术化方向的研究重点[21-25]；四是结构化语库能否稳定提升学生辨证论治能力、处方分析能力和真实病例复盘能力，仍需在后续教学实践中通过对照研究、形成性评价和访谈反馈进一步验证。

4 结语

本研究围绕中医门诊临床经验传承与教学数字化的现实需求，构建完整的NLP与信息抽取技术流程，从而实现从非结构化中医脾胃病门诊病案到可查询、可统计的结构化医案语库的转换，进一步实现“症状—证型—方药”及纵向病程的标准化存储。研究不仅为中医病种谱分析及临床规律挖掘提供了数据支撑，更为丰富中医CBL教学资源、展示真实世界动态辨证思维提供了创新的数字化教具库，对促进中医临床教学的智能化发展具有重要实践意义。

参考文献：

[1] 柴华，路海明，刘清晨. 中医自然语言处理研究方法综述[J]. 医学信息学杂志，2015，36(10)：58-63.

[2] 包小源，黄婉晶，张凯，等. 非结构化电子病历中信息抽取的定制化方法[J]. 北京大学学报(医学版)，2018，50(2)：256-263.

[3] 王维笑，费晓璐，闾海荣，等. 基于自然语言处理与结构化算法的病历信息高精度抽取方法研究[J]. 中国数字医学，2024，19(5)：40-48.

[4] Meystre S M，Savova G K，Kipper-Schuler K C，et al. Extracting information from textual documents in the electronic health record：a review of recent research[J]. Yearb Med Inform，2008：128-144.

[5] Wang Y，Wang L，Rastegar-Mojarad M，et al. Clinical information extraction applications：a literature review[J]. J Biomed Inform，2018，77：34-49.

[6] Fu S，Chen D，He H，et al. Clinical concept extraction：a methodology review[J]. J Biomed Inform，2020，109：103526.

[7] Yao L，Jin Z，Mao C，et al. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora[J]. J Am Med Inform Assoc，2019，26(12)：1632-1636.

[8] Chen L，Song L，Shao Y，et al. Using natural language processing to extract clinically useful information from Chinese electronic medical records[J]. Int J Med Inform，2019，124：6-12.

[9] Zhang Y，Wang X，Hou Z，et al. Clinical named entity recognition from Chinese electronic health records via machine learning methods[J]. JMIR Med Inform，2018，6(4)：e50.

[10] Li Y，Wang X，Hui L，et al. Chinese clinical named entity recognition in electronic medical records：development of a lattice long short-term memory model with contextualized character representations[J]. JMIR Med Inform，2020，8(9)：e19848.

[11] Chen P，Zhang M，Yu X，et al. Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT[J]. BMC Med Inform Decis Mak，2022，22(1)：315.

[12] Wang W，Li X，Ren H，et al. Chinese clinical named entity recognition from electronic medical records based on multisemantic features by using the RoBERTa-wwm-ext，IDCNN，and CRF models[J]. JMIR Med Inform，2023，11：e44597.

[13] An Y，Xia X，Chen X，et al. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF[J]. Artif Intell Med，2022，127：102282.

[14] Wu Y，Jiang M，Lei J，et al. Named entity recognition in Chinese clinical text using deep neural network[J]. Stud

Health Technol Inform,2015,216:624-628.
[15] Li Y, Du G, Xiang Y, et al. Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge [J]. J Biomed Inform, 2020, 106:103435.
[16] Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records [J]. NPJ Digit Med, 2022,5(1):194.
[17] Thistlethwaite JE, Davies D, Ekeocha S, et al. The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide No. 23 [J]. Med Teach,2012,34(6):e421-e444.
[18] McLean SF. Case-based learning and its application in medical and health-care fields: a review of worldwide literature [J]. J Med Educ Curric Dev, 2016, 3: JMECD. S20377.
[19] Korniichuk OY, Bambyzov LM, Kosenko VM, et al. Application of the case study method in medical education [J]. Int J Learn Teach Educ Res,2021,20(7):175-191.
[20] Tsekhmister Y. Effectiveness of case-based learning in medical and pharmacy education: a meta-analysis [J]. Electron J Gen Med,2023,20(5):em515.
[21] Wong TH, Ip EJ, Lopes I, et al. Pharmacy students' performance and perceptions in a flipped teaching pilot on cardiac arrhythmias [J]. Am J Pharm Educ, 2014, 78 (10):185.
[22] Gu B, Shao V, Liao Z, et al. Scalable information extraction from free text electronic health records using large language models [J]. BMC Med Res Methodol, 2025, 25(1):23.
[23] Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation [J]. BMJ Health Care Inform,2025,32(1):e101139.
[24] Duan Y, Qiu J, Yang R, et al. Research on a traditional Chinese medicine case-based question-answering system integrating large language models and knowledge graphs [J]. Front Med (Lausanne),2025,11:1512329.
[25] Yang A, Li A, Yang B, et al. Qwen3 technical report [EB/OL]. (2025-05-14) [2026-05-01]. https://arxiv. org/obs/2505. 09388.

(收稿日期:2026-05-25)