

基于本地大模型的骨质疏松专病数据库构建流程优化研究^{*}

王耀国¹ 唐诗诗² 刘泓泽² 安雨婷¹ 周 毅²

(¹ 新疆医科大学公共卫生学院 乌鲁木齐 830017 ² 中山大学中山医学院 广州 510080)

〔摘要〕 目的/意义 利用本地大模型和知识库优化骨质疏松专病数据库构建流程，降低医疗数据泄漏风险。方法/过程 收集相关临床研究文献和业务数据信息，构建本地知识库；基于本地大模型应用架构，使用本地知识库限定本地大模型的生成内容；基于本地大模型应用架构图形问答界面和编程接口，通过自然语言问询和编程引导本地大模型生成骨质疏松专病数据库字段集。结果/结论 本地大模型结合本地知识库可以总结骨质疏松既往研究中涉及的字段，有效提高骨质疏松专病数据库构建效率和质量，但处理结果需要人工核查。

〔关键词〕 专病数据库；大模型；本地化部署；自然语言理解；非结构化数据处理

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2025.01.013

Study on Optimization of Osteoporosis Specialized Disease Database Construction Process Based on Local Large Models

WANG Yaoguo¹, TANG Shishi², LIU Hongze², AN Yuting¹, ZHOU Yi²

¹School of Public Health, Xinjiang Medical University, Urumqi 830017, China; ²Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China

〔Abstract〕 **Purpose/Significance** To utilize local large models combined with local knowledge base to optimize the osteoporosis disease database construction process, and to reduce the risk of medical data leakage. **Method/Process** Firstly, relevant clinical research literature and business data information are collected and a local knowledge base is built. Then, based on the local large model application architecture, the local knowledge base is used to limit the generated content of the local large model. Finally, based on the graphical question-and-answer interface and programming interface of the local large model application architecture, natural language queries and programming methods are used to guide the local large model to generate osteoporosis disease database field sets and data production. **Result/Conclusion** The local large models combined with the local knowledge base can summarize the fields involved in previous osteoporosis researches and effectively improve the efficiency and quality of osteoporosis disease database construction, but the processing results need to be manually verified.

〔Keywords〕 specialized disease database; large model; localized deployment; natural language understanding (NLU); unstructured data processing

〔修回日期〕 2024-09-24

〔作者简介〕 王耀国，硕士，发表论文 3 篇；通信作者：周毅，博士，教授，博士生导师。

〔基金项目〕 国家重点研发计划（项目编号：2022YFC3601600）；广东省自然科学基金项目（项目编号：2024A1515011989）；新疆维吾尔自治区人民医院院内课题（项目编号：20230117）。

1 引言

骨质疏松是一种骨量减少和微结构退化导致脆性增加和骨折风险增加的疾病。随着医疗信息化的发展,用于临床研究的数据种类和数量日益增多,能否根据研究目的高效收集、整理临床数据成为影响骨质疏松临床研究效率和质量的重要因素^[1-2]。骨质疏松专病数据库用于存储、管理和分析骨质疏松相关数据^[3-4],可为临床研究提供有效支持。骨质疏松临床数据的分析具有一定复杂性,其多源异构临床数据完成集成后还需要清洗才可以进行分析并用于临床研究^[5-6]。

主流专病数据库构建流程为,先提出研究假设,再确定疾病相关字段集,然后进行临床数据抽取或生产。目前疾病相关字段集的确定方案主要为基于研究目的和临床经验,从疾病临床诊疗指南、已发表相关文献中确定初始字段集合,该研究方案依赖于研究者的临床经验。近年来,随着人工智能技术的发展,大模型在自然语言理解和生成方面取得显著进展,能够根据所输入文本生成相应结果^[7-9]。利用大模型优化专病数据库构建流程具有一定可行性和潜力^[10]。从大数据分析和临床数据挖掘角度来看,可以利用大模型的归纳、总结能力,处理临床数据中的长文本字段,全面提取医疗事件类别,然后进行筛查,作为疾病相关字段集的补充。

2 研究目的

利用本地大模型、本地知识库总结文档,通过给定明确的本地系统架构上下文信息辅助生成可直接运行的代码。基于本地大模型的自然语言理解和文档提取、总结能力,从既往骨质疏松临床研究文献中总结提取相关研究字段集,经临床医师团队二次核查和讨论,作为骨质疏松专病字段集的补充和完善方案。在数据不离开医院的前提下,基于本地化部署大模型,提高对骨质疏松临床非结构化数据的结构化解析效率和质量,如对电子版骨密度检查报告中表格数值的提取和结构化。同时,专病数据

库构造语句繁多且数据来源复杂,工作量大,每个字段需要考虑兼容性问题。利用大模型的代码生成能力^[8,11-13],基于医院后台数据架构信息和疾病信息构建向量数据库,生成相关编程语句,优化专病数据库构建流程,提高构建效率和数据质量^[14-15]。

3 系统架构

在专病数据库构建各环节使用不同本地大模型进行流程优化。骨质疏松专病数据库构建流程大致分为 3 个阶段。一是根据研究假设,确定要收集的临床数据字段集。二是数据抽取,基于抽取、转换和加载(extraction - transformation - loading, ETL)技术与临床数据字段集,确定医院相关系统的底层数据位置,抽取的同时完成多源异构数据集成。三是数据清洗,对来源于不同系统的临床数据统一处理格式、单位、值域等,剔除不符合逻辑的值,根据研究假设和数据量处理缺失值、离群值;同时对非结构化数据进行结构化处理,抓取目标字段。本研究基于本地大模型对上述流程进行优化。第 1 阶段,专病数据库字段集确定。基于骨质疏松临床研究文献构建本地知识库,利于本地大模型总结既往研究文献中出现的临床字段,并与本单位临床数据对照,形成骨质疏松专病数据库字段集。第 2 阶段,数据抽取。收集业务数据并整理为业务文档,构建指标平台,在利用大模型生成代码时,通过相似性检索给定本地大模型明确的本地系统架构信息,使其生成的代码细节中能够包含真实生产环境的上下文信息,例如,筛选能够参与专病库构建的数据表,每张数据表中包含的字段及其含义;通过多轮问询,引导大模型生成符合实验环境的数据抽取语句(一般为结构化查询语言)。第 3 阶段,数据清洗。使用多模态大模型提高多模态数据的处理效率和质量。目前专病数据库主要针对关系型数据进行集成和处理,通过结构化查询语言直接从原始数据中抓取即可,对于需要从“一诉五史”等长文本中二次生产的字段,可以通过调用本地大模型接口进行自然语言处理(natural language processing, NLP)。除此之外,对文件、图像类临床报告也需进行处理。处理方案是先将其统一转化为图片,再将图片输入大

模型进行光学字符识别 (optical character recognition, OCR), 提取其中的文本和数值, 最后进行格式化。

4 研究内容

4.1 大模型本地化部署

为确保信息安全, 用于科研的医疗数据需经数据脱敏等一系列复杂处理, 且必须在特定的物理位置处理。大模型本地化部署有助于减少对第三方服务提供商的依赖, 提高医疗机构数据资产自主可控性。目前, 开源大模型种类繁多, 为解决各种任务还需自定义模型。因此在专病数据库构建的不同环节采用不同大模型, 例如, 通过相关文献总结确定骨质疏松专病数据库字段集时采用 Qwen7b 大模型, 在数据处理阶段采用 GLM-4V-9B 多模态大模型。多种类型大模型部署步骤较为烦琐^[16], 为简化大模型本地化部署流程且更高效应用不同类型的大模型解决专病数据库构建流程中遇到的问题, 本研究基于开源大语言模型框架 Ollama, 通过将模型权重、配置和数据打包成单一程序包, 优化参数设置和显卡、内存等资源配置细节, 实现大模型更灵活的本地化部署, 以适应专病数据库构建需求。

4.2 构建本地知识库

基于向量数据库构建本地知识库, 构建流程, 见图 1。首先, 收集骨质疏松相关文献信息, 构建长文本集合。目前涉及资料主要为文本数据, 如疾病临床诊疗指南、研究论著等。其次, 上传相关文献到指定存储位置, 基于自然语言处理算法对文本嵌入处理, 即对长文本编码, 构建 $m \times n$ 的向量堆栈, 进行分片处理以保障后期查询性能。最后, 通过向量嵌入方式存储在向量数据库中, 形成本地知识库。本地知识库具有查询功能, 用于处理大模型查询请求。具体流程为基于检索增强生成技术将向量数据库和大模型问答能力相结合。当用户在图形界面对大模型提出问题时, 通过大模型应用框架将用户输入查询转发给指定大模型, 将输入语句 (一般为自然语言) 转化为向量形式, 并通过大模型应

用框架转发给本地知识库查找向量数据库中相关部分, 通过计算余弦距离得出相似度, 完成相似性检索, 最后返回本地知识库中最相似结果用于指导大模型最终结果的生成。

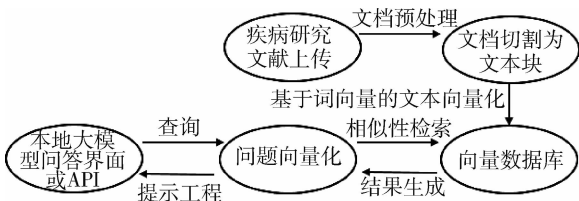


图 1 本地知识库构建流程

4.3 基于业务文档指导代码生成

大模型具备一定代码生成能力, 但是因为缺少实际环境信息, 其生成结果只能作为参考, 需要修改后使用。在此基础上, 收集本地业务知识相关数据, 如医院后台架构信息以及各种字段的解释说明等文档, 形成业务文档。在利用大模型生成代码时, 给定本地大模型明确的系统架构信息, 使其生成的代码细节中能够包含真实生产环境的上下文信息。

4.4 基于大模型的非结构化数据处理

利用医院平台已有无纸化系统, 批量获取骨密度报告的 PDF 文件, 解析后提取目标字段。分析发现 PDF 报告大致分为两类, 一类支持读写, 可以通过简单代码提取; 另一类由图片生成的 PDF 文件则无法直接提取, 需要进行 OCR 处理。但由于骨密度报告结构较为复杂, 传统 OCR 工具识别一般会丢失原逻辑结构。在对比传统 OCR 工具和基于深度学习的 OCR 工具后, 基于 GLM-4V-9B 多模态大模型, 先将第 2 类 PDF 报告文件剪切成多张图片, 然后依次进行解析, 提取图片中的文字和表格, 并将结果制作为 JSON 格式, 通过键名获取目标数值并存储到骨质疏松专病数据库中。

5 研究结果

5.1 辅助生成骨质疏松专病字段集

目前专病数据库字段集主要是由研究者结合主观

经验、临床客观存在且可以使用的数据大致确定。该方法受限于研究者经验，以及研究文献总结提取能力、记忆力，导致专病数据库生产完成后或数据分析时可能需要增加专病数据库字段，重新生产。

为了得出大模型对流程优化的量化指标，以骨质疏松筛查为主题，收集中国知网和谷歌学术等网站 2021—2023 年相关论著、指南等文献，共计 200 余篇。分别采用传统人工阅读方式与本地大模型方式，比较骨质疏松专病字段集构建所花费的时间。采用人工方式从 200 余篇骨质疏松相关文献中总结研究字段，耗时 1 ~ 2 周。采用大模型方式，基于

本地知识库耗时一天完成 200 余篇文献骨质疏松筛查相关字段总结。具体流程为，将相关文献上传到本地知识库中，采用大模型进行文本向量化处理，总结文献中与目标疾病相关的临床字段，进行向量相似性检索，从既往骨质疏松研究文献中总结提取相关特征。之后基于本地知识库，利用本地大模型框架实现以聊天形式检索文档，并进行总结。基于 AnythingLLM 大模型构建的本地知识库支持展示引文，实现知识溯源。与本地临床数据字段名称对照，获得用于生产的字段集，作为待研究的骨质疏松专病字段集补充和参考，见表 1。

表 1 骨质疏松专病数据库字段集（部分）

模块名称	子模块名称	字段名称	数据处理
患者人口学信息	主索引、居住地	患者编号、性别、年龄、职业、现住址、是否死亡、死亡时间等	直接抽取
就诊记录	门诊信息、住院信息	患者编号、门诊日期、门诊诊断、入院日期、出院诊断、出院日期等	直接抽取
病历信息	生命体征、病程记录	骨痛、乏力、脊柱变形、身材缩短、骨折、恐惧、焦虑、抑郁、药物应用史、家族史、既往史等	直接抽取 + NLP 处理
检查信息	超声检查、双能 X 线骨密度检查	骨密度测定、胸椎、腰椎 X 线侧位片等	NLP 处理 + OCR 处理
检验信息	生化指标	血清碱性磷酸酶、骨转换标志物、骨钙素、血清 I 型原胶原 C - 端前肽等	直接抽取
医嘱信息	药品医嘱	用药时长、用药名称、用药剂量	直接抽取 + NLP 处理
手术信息	手术类型	手术名称、手术部位、手术日期等	直接抽取 + NLP 处理

5.2 非结构化数据的结构化处理

本研究中待处理的非结构化数据主要为“一诉五史”等长文本信息和检查报告。以骨密度检查报告为例，2021—2024 年骨密度检查报告共计 8 万余份 PDF 文件。选用可本地化部署的 GLM - 4V - 9B 多模态大模型，在保证数据不离开医院，避免数据泄漏风险的同时也对非结构化数据进行了高效处理。从图片中提取指定表格信息的速度依赖于硬件条件，本研究中本地部署传统 OCR 工具并识别一份骨密度报告的时间为 7 ~ 15 秒，基于本地大模型处理一份报告约需 20 秒。但是识别结果差异明显，传统的能够用于本地部署的 OCR 工具，其识别结果大多会丢失逻辑结构，解析结果不能直接用于后续研究，需要二次编程提取相关数据并保存。而基于本地多模态大模型的处理

结果可以指定数据结构输出，结果可直接存入数据库，无须二次处理，见表 2。

表 2 本地多模态大模型与传统 OCR 工具对比

对比项	传统 OCR 工具	本地多模态大模型
数据是否保留原有表格结构	否	是
数据是否离开本地	是	否
支持多种数据格式	是	是

6 结语

本研究采用 Qwen7b 和 GLM - 4V - 9B 大模型优化骨质疏松专病数据库构建流程。通过实验验证发现，利用本地大模型基于骨质疏松研究文献构建的本地知识库，可以总结骨质疏松临床研究字段，作

为构建骨质疏松专病数据库字段集的补充。通过收集业务知识构建本地指标文档,通过提示工程引导本地大模型初步生成具备实验环境信息的编程语句,提高了专病数据库的构建效率。医疗数据处理有严格的信息安全要求,最好能在医院内部进行。基于本地多模态大模型处理非结构化数据既能保证数据安全,又能够提高生成效率和质量。本研究采用的方案尚存在不足:由于本地知识库的相关文献数量较少,本地大模型生成的用于参考和补充的专病字段集也较少,未来将进一步收集相关疾病文献,扩展本地知识库规模,增大专病字段集。由于医疗数据的复杂性和多样性^[17-18],生成结果仍然需要人工校对和调整,未来研究可进一步探索提高大模型在专病数据库构建中的准确性和可靠性,以满足医疗信息化需求。

作者贡献: 王耀国负责研究设计、论文撰写;唐诗诗负责论文撰写;刘泓泽负责实验方案实施;安雨婷负责文献分析、论文撰写;周毅负责提供指导、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 李榕,潘常青,袁骏毅,等. 基于大数据技术的智能化房颤专病数据库建设和应用 [J]. 中国数字医学, 2022, 17 (2): 95-100.
- 2 薛万国,乔岫,车贺宾,等. 临床科研数据库系统的现状与未来 [J]. 中国数字医学, 2021, 16 (1): 2-6.
- 3 龙思哲,吴震天,黎鹏安,等. 基于数据治理的专病数据库建设实践 [J]. 医学信息学杂志, 2022, 43 (7): 20-25.
- 4 刘迷迷,杜国霞,周毅,等. 专病数据库建设与应用研究 [J]. 医学信息学杂志, 2021, 42 (11): 81-86, 93.
- 5 张弘政,刘迷迷,李琳,等. 基于通用数据模型的健康医疗大数据平台数据治理研究 [J]. 医学信息学杂志, 2022, 43 (6): 2-7, 13.
- 6 LIU M, LUO J, LI L, et al. Design and development of a disease-specific clinical database system to increase the availability of hospital data in China [J]. Health information

science and systems, 2023, 11 (1): 11.

- 7 缪青海,王兴霞,杨静,等. 从基础智能到通用智能: 基于大模型的 GenAI 和 AGI 之现状与展望 [J]. 自动化学报, 2024, 50 (4): 674-687.
- 8 KATSOGIANNIS - MEIMARAKIS G, KOUTRIKA G. A survey on deep learning approaches for text-to-SQL [J]. The VLDB journal, 2023, 32 (4): 905-936.
- 9 CHANG Y, WANG X, WANG J, et al. A survey on evaluation of large language models [J]. ACM transactions on intelligent systems and technology, 2024, 15 (3): 1-45.
- 10 胡振生,杨瑞,朱嘉豪,等. 大语言模型在医学领域的研究与应用发展 [J]. 人工智能, 2023, (4): 10-19.
- 11 KASNECI E, SEBLER K, KÜCHEMANN S, et al. ChatGPT for good? On opportunities and challenges of large language models for education [J]. Learning and individual differences, 2023, 103: 102274.
- 12 SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620 (7972): 172-180.
- 13 ZHONG V, XIONG C, SOCHER R. Seq2sql: generating structured queries from natural language using reinforcement learning [EB/OL]. [2024-01-08]. <https://arxiv.org/pdf/1709.00103>.
- 14 BAIG M S, IMRAN A, YASIN A U, et al. Natural language to sql queries: a review [J]. International journal of innovations in science technology, 2022, 4 (1): 147-162.
- 15 YANG R, TAN T F, LU W, et al. Large language models in health care: development, applications, and challenges [J]. Health care science, 2023, 2 (4): 255-263.
- 16 KAPLAN S J. Designing a portable natural language database query system [J]. ACM transactions on database systems, 1984, 9 (1): 1-19.
- 17 SUBRAMANIAN C R, YANG D A, KHANNA R. Enhancing health care communication with large language models—the role, challenges, and future directions [J]. JAMA network open, 2024, 7 (3): e240347.
- 18 TRUMMER I. CodexDB: synthesizing code for query processing from natural language instructions using GPT-3 codex [J]. Proceedings of the VLDB endowment, 2022, 15 (11): 2921-2928.