基于 NLP 构建病历后结构化专病数据库探索与实践*

张亚男[1] 董 亮[1] 何 萍[2]

([1] 上海中医药大学附属龙华医院 上海 200032 [2] 上海申康医院发展中心 上海 200041)

〔摘要〕 **目的/意义** 建设基于结构化电子病历的专病数据库，提高专病数据库规范性和可用性，提高临床科研工作效率。**方法/过程** 采用模板化输入、自然语言处理等技术，将非结构化电子病历转化为结构化电子病历，基于结构化电子病历构建专病数据库。**结果/结论** 龙华医院基于结构化电子病历建设的银屑病专病数据库分中心，为临床科研人员提供结构化科研数据源，辅助提升分析效率；同时有效支撑上海申康“基于多中心的银屑病专病大数据临床科研随访一体化平台”建设，有助于专病数据库高质量、规模化发展。

〔关键词〕 自然语言处理；结构化电子病历；专病数据库

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.09.013

Exploration and Practice of Constructing a Structured Specialized Disease Database Based on NLP for Medical Records

ZHANG Yanan[1], *DONG Liang*[1], *HE Ping*[2]

[1] *Longhua Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai 200032, China*; [2] *Shanghai Hospital Development Center, Shanghai 200041, China*

〔**Abstract**〕 **Purpose/Significance** To construct a specialized disease database based on structured electronic medical records (EMR), and to improve the standardization and usability of the specialized disease database and improve the efficiency of clinical scientific research. **Method/Process** By using templated input, natural language processing (NLP) and other technologies, unstructured EMR are converted into structured EMR, and a specialized disease database is constructed based on structured EMR. **Result/Conclusion** The psoriasis specialized disease database sub center built by Longhua Hospital based on structured EMR provides clinical researchers with a structured scientific research data source, assists researchers in improving analysis efficiency, and effectively supports the construction of Shanghai Shenkang's "multi center psoriasis specialized disease big data clinical research follow-up integrated platform", which is conducive to the high-quality and large-scale development of specialized disease databases.

〔**Keywords**〕 natural language processing (NLP); structured electronic medical records (EMR); specialized disease database

1 引言

随着医疗信息化的发展，医生普遍使用电子病历系统记录患者的病情及诊疗方案，这些数据具备较高的临床科研价值。受诊疗方案繁杂且不同地区

〔修回日期〕 2024-05-18

〔作者简介〕 张亚男，高级工程师，发表论文 5 篇。

〔基金项目〕 上海申康医院发展中心管理研究项目（项目编号：2023SKMR-22）。

医生语言表述习惯不同等因素的影响，电子病历系统普遍存在数据格式多样化、表述不统一等问题，给临床科研人员筛选和分析带来困难。因此，要提高病历数据的可用性，首先要对病历结构化处理。

目前，电子病历结构化处理的基本手段是使用模板化输入生成半结构化电子病历[1]，但是半结构化电子病历中仍包含非结构化内容。为进一步实现电子病历结构化，医疗信息行业探索使用自然语言处理（natural language processing，NLP）技术，经过自然语义理解、命名实体识别[2]、关系抽取[3]、语义分析等步骤提高病历数据的结构化程度[4]。本文介绍龙华医院基于 NLP 对电子病历后结构化处理，并依托结构化电子病历构建银屑病专病数据库的过程，探讨 NLP 技术提升电子病历数据规范化水平、辅助科研人员高效高质量完成临床科研工作的技术路径及效果。

2 专病数据库建设可行性分析

20 世纪 60 年代欧美国家开始研究 NLP 技术在临床医学领域的应用；到 20 世纪 90 年代，大量医学知识库如医学系统命名法 - 临床术语（systematized nomenclature of medicine clinical terms，SNOMED CT）建立起来，这些知识库因集成了大量临床常用的医学诊断、检查、治疗术语而被广泛使用[5]；近 10 年来，深度学习等技术开始应用于电子病历场景，例如由谷歌公司、斯坦福大学等组成的团队开发了基于深度学习的 NLP 方法[6]，处理 216 000 条电子健康记录，用于预测患者疾病发病时间和死亡率[7]。

从目前的研究成果可知，NLP 技术在处理英文电子病历方面取得了部分成效，但处理中文电子病历时仍面临分词难度大、实体识别困难、语义理解复杂等难题。中文电子病历包含大量专病医学术语、缩写，且相同病情的表述习惯也因地区、医院和医生的不同而呈现多样性。因此，目前 NLP 技术在国内医疗行业的应用范围仍局限于结构化程度较高的病历内容，如商汤公司自主研发的 SenseCare 智慧诊疗平台侧重于影像报告领域应用研究[8]。

针对病情表述多样性问题，笔者所在研究团队经过深入调研和分析发现：上海地区的银屑病专科病种，统一由上海市皮肤科临床质量控制中心管理，遵循相同的质控标准[9]，使用相近的医学术语库，病历格式和表述趋同性较强。基于结构化电子病历的银屑病专病数据库[10]仅在上海医联体[11]内共享应用，可解决病情表述多样性的问题，保证病历数据在多中心间的共享性和可用性。

3 基于 NLP 技术的银屑病专病数据库建设

3.1 系统架构

龙华医院在标准规范体系和安全防护体系的支撑下，使用 NLP 技术对银屑病电子病历后结构化处理，搭建龙华医院银屑病专病数据库分中心[12]，并与上海申康“基于多中心的银屑病专病大数据临床科研随访一体化平台”对接，系统架构，见图 1。

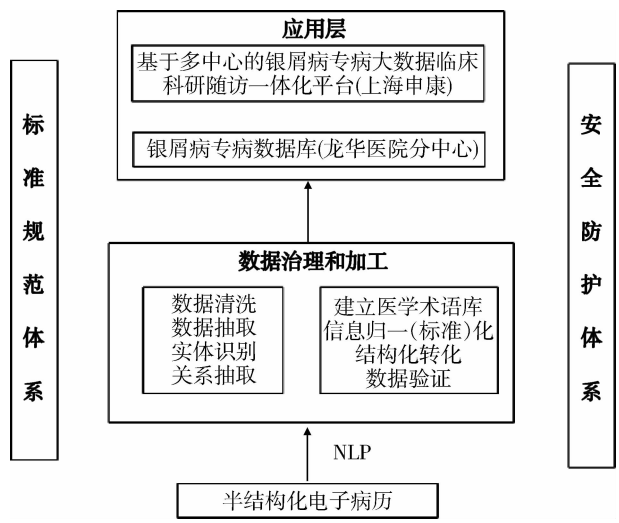

图 1 银屑病专病数据库架构

3.2 技术路线

构建龙华医院银屑病专病数据库分中心的核心工作是使用 NLP 技术实现电子病历的后结构化。本系统探索使用多路召回、Rank 匹配计算、个数预测等 NLP 技术，由模板化输入、数据清洗、数据抽取、实体识别、关系抽取、建立医学术语库、信息归一（标准）化、结构化转化、数据验证 9 个步骤

组成电子病历后结构化技术路线[13-17]。

3.2.1 模板化输入 选取银屑病诊疗信息密度最高[18]的 4 份半结构化电子病历模板：门诊初诊模板、门诊复诊模板、入院记录模板、出院小结模板。以出院小结模板为例，包含皮疹部位、分布、皮肤原发损害等结构化模块，但模块内容仍为文本格式，须使用 NLP 技术进行后结构化处理[19]。

3.2.2 数据清洗 主要包括删除与治疗方案相关性低且缺失严重的数据、自动补全可根据上下文推断出的缺失数据、由医疗专家人工补足或确认重要且无法自动处理的数据等操作。

3.2.3 数据抽取 抽取 505 例银屑病患者的门诊初诊、门诊复诊、入院记录、出院小结病历数据。

3.2.4 实体识别 采用 BIO 标注方式，人工对银屑病诊疗过程涵盖的 41 个实体逐项处理，将每个实体的第一个字符标签设为 B，剩余字符标签设为 I，将标注好的数据输入命名实体识别模型进行训练，以得到高质量的命名实体识别模型。

3.2.5 关系抽取 从文本格式的病历中抽取（主体、关系、客体）三元组，建立实体之间的关系，如“张三于 2019 年连续高强度工作 3 月后，在 XX 医院确诊银屑病”，通过关系抽取操作，得到实体关系三元组实例（患者、发病诱因、连续高强度工作）[20]。

3.2.6 建立医学术语库 医学术语库是电子病历后结构化的基础，医学术语库的权威程度，影响结构化电子病历的专业程度。在银屑病医学术语库的建设过程中，中医诊断参考《中医病证分类与代码》（GB/T 15657—2021）等 4 份国家标准，西医诊断参考《疾病和有关健康问题的国际统计分类》（ICD-10），手术诊断参考《国际疾病分类第九版临床修订版手术与操作》（ICD9-CM-3），并结合中华医学会皮肤性病学分会银屑病专业委员会制定的《中国银屑病诊疗指南》和上海市皮肤科临床质量控制中心质控指标，设计银屑病标准数据集及其数据元，并建立银屑病值域字典，见表 1。

表 1 银屑病数据元值域字典（部分）

数据元编码	数据元名称	值域
yxbfx	银屑病分型	01：关节型银屑病
		02：脓疱型银屑病
		03：寻常型银屑病
jbzlqk	疾病治疗情况	01：手术治疗
		02：未给予治疗
		03：药物治疗
jbhjjt	疾病缓解状态	01：好转
		02：稳定
		03：加重

3.2.7 信息归一（标准）化和结构化转化 电子病历后结构化过程中两步密切相关的核心操作。（1）信息归一（标准）化。系统依据银屑病医学术语库，使用多路召回、Rank 匹配计算、个数预测等技术，实现实体与数据元值域字典表的映射，将相同的数据元使用统一的词汇表述。多路召回通过“标准词”查询实现“映射原词”与“标准词”的映射，“标准词”查询分为 3 类：将“映射原词”与“标准词”词库进行相似度计算，将“映射原词”与历史训练集中的“映射原词”进行相似度计算，将“映射原词”与当前查询中的“映射原词”匹配。匹配计算基于“双向编码器表征量”“映射原词”和“标准词”等基本信息，首先增加两个池化层，进行 0 和 1 分类；然后通过相似度计算选取与“标准词”相似度排名末 20 位的“映射原词”，完成负样本构造；最后增加正样本数量，经过“映射原词”与“标准词”相似度计算后再匹配。个数预测基于“双向编码器表征量”构建 3 类“映射原词”与“标准词”的数量对应关系，数量对应关系分为 0~1 个、1~2 个、大于 2 个，当数量对应关系预测≤2 时，返回 top-k，当数量对应关系预测>2 时，按照得分排序，选择阈值>0.5 的“映射原词”与“标准词”匹配，解决“映射原词”与“标准词”多对一的问题。（2）结构化转化。其基

础操作是根据医疗专家设计的诊疗指南和质控指标设计结构化电子病历数据库表及字段，提取经过信息归一（标准）化处理后的明确症状、伴有疾病等实体进行结构化转化操作，最后将转化后生成的实体以数据集形式存储在结构化电子病历数据库中。

以“现病史”结构化电子病历转化过程为例。首先将文本格式数据中的“伴轻度瘙痒”存入“自觉症状”字段，“无明显诱因”存入“初次发病诱因”字段，“四肢红色鳞屑性丘疹”存入“初次发病部位”字段，完成实体向数据元的结构化转化操作；然后依据“建立医学术语库”步骤中设计的银屑病标准数据集和数据元对应关系，使用聚类匹配等技术，将数据元聚类形成数据集。根据医疗专家设计的诊疗指南和质控指标，设计结构化电子病历数据库表结构，以数据集形式存储归一化处理后的结构化电子病历实体信息，形成数据元级别的结构化电子病历数据[21]，见图 2。

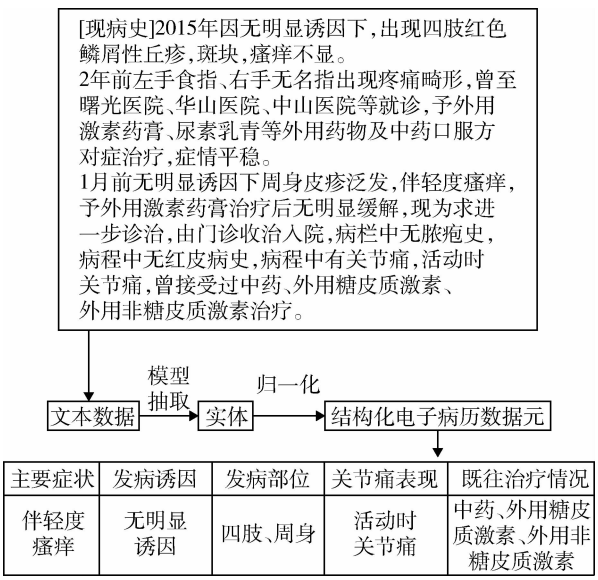

图 2 文本数据的结构化转化过程

3.2.8 数据验证 基于 BIO 三元标注格式[22]将字符标签设为 B 和 I 的实体认定为正例，字符标签设为 O 的实体认定为负例，使用精确率、召回率和 *F*1 分数进行实体识别数据验证。其中，高精确率意味着模型很少将非实体错误地识别为实体，高召回率意味着模型很少错过正例识别，*F*1 用于平衡精确率和召回率。抽取 87 份“现病史”后结构化电子病历，得出主诉中的银屑病病程实体精确率为 69.48%，召回率为 71.25%，*F*1 为 70.35%；自觉症状实体精确率为 73.96%，召回率为 70.58%，*F*1 为 72.23%；41 个实体的平均精确率为 80.00%、召回率为 76.49%，*F*1 为 78.12%，满足电子病历后结构化的基本质量要求。

4 应用成效

基于 NLP 技术生成的银屑病专病数据库具备自动化数据处理功能，可替代临床医生处理繁复的科研病历记录和整理工作，提高工作效率。专病库面向 2020 年 12 月—2022 年 9 月的 505 例银屑病门诊患者和 165 例住院患者，自动生成 554 份结构化电子病历，其中 180 份门诊初诊、215 份门诊复诊、109 份入院记录、50 份出院小结[23]，涵盖 389 项数据元，相较于以往手工生成专病库，时长缩短 70% 以上。在 2022 年 7 月 29 日发布的质控报告中，龙华医院银屑病专病数据库数据元完整性 100%，规范性 98.85%，一致性 100%，表明该专病库具备较高的数据质量。龙华医院作为上海申康“基于多中心的银屑病专病大数据临床科研随访一体化平台”[24]项目的分中心单位，完成了银屑病中医类数据规范（数据元值域和数据集）的制定工作，为银屑病医联体的专病数据库增补了中医药治疗方案，见图 2[25]。

表 2 中医诊断代码和名称（部分）

中医诊断编码	中医诊断名称	中医证候名称
A08.01.07	湿疮	八纲证候类
A08.01.15	白疕	八纲证候类
A08.01.15	白疕	肺经风热证
A08.01.20	粉刺	肺经风热证
A08.01.40	白驳风	八纲证候类

5 结语

本文主要介绍了一种基于 NLP 技术构建病历后结构化专病数据库的技术路径。以结构化电子病历

为核心的专病数据库可有效支撑医疗专家精确检索患者全流程诊疗信息，深入探索疾病的发病机制、病程发展和治疗效果，开展临床决策和医学研究，对于提高医疗水平、改善患者治疗效果和推动医学研究具有重大意义。但是由于各专病个性化强、专业术语交叉率低等特点，本研究只完成了一种专病的电子病历后结构化处理进而构建专病数据库，暂未解决针对多个专病普遍适用的病历后结构化处理问题。未来，医疗信息行业可引入深度学习等人工智能技术，开展大样本诊疗病历数据的分析和学习，提取更多的疾病相关特征和知识，完善疾病的描述和分类，拓展疾病领域，探索更多创新性研究和应用，推动专病数据库建设的智能化、普适化发展。

作者贡献：张亚男负责论文撰写与修订；董亮负责提出研究思路；何萍负责提供指导。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 马欢欢，孔繁之，高建强．中文电子病历命名实体识别方法研究［J］．医学信息学杂志，2020，41（4）：24－29.

2 汤昊成，苏万春，冀秀元，等．面向淋巴水肿疾病的电子病历命名实体识别应用研究［J］．医学信息学杂志，2024，45（2）：52－58.

3 张维宁，申喜凤，李美婷，等．融合关系标签和位置信息的中文医疗文本因果关系抽取方法研究［J］．医学信息学杂志，2024，45（1）：21－26.

4 吴宗友，白昆龙，杨林蕊，等．电子病历文本挖掘研究综述［J］．计算机研究与发展，2021，58（3）：513－527.

5 刘盛宇，胡拯涌，段一凡，等．欧洲 ELIXIR 生物医学数据工具服务平台体系实践与启示［J］．医学信息学杂志，2023，44（11）：63－70.

6 肖仰华，徐一丹．大规模生成式语言模型在医疗领域的应用：机遇与挑战［J］．医学信息学杂志，2023，44（9）：1－11.

7 任媛渊，丁福，付荣娟．国外基于机器学习的住院患者跌倒风险预测模型构建研究及其启示［J］．医学信息学杂志，2023，44（10）：63－67，80.

8 龚宇新，向菲，应葵．医学影像与自然语言处理多模态探索研究［J］．医学信息学杂志，2024，45（1）：33－38.

9 刘伟伟，王立军，庞娟，等．基于自然语言处理的肿瘤专科病历质控系统建设［J］．医学信息学杂志，2024，45（2）：77－81.

10 石晶金，袁瑞，冯雨萱，等．跨区域专科联盟数据共享建设现状、影响因素与策略分析［J］．医学信息学杂志，2023，44（11）：30－34.

11 萧锘，叶琪，刘传丰，等．区域临床科研大数据平台设计与实现［J］．中国卫生信息管理杂志，2022，19（5）：673－680.

12 袁骏毅，潘常青，李榕，等．基于临床数据中心的冠心病专病数据库的构建与实现［J］．中国卫生信息管理杂志，2022，19（5）：707－712.

13 胡建平，张晓祥，庹兵兵，等．中文医学术语标准体现构建研究［J］．中国卫生信息管理杂志，2023，20（1）：19－24.

14 崔博文，金涛，王建民．自由文本电子病历信息抽取综述［J］．计算机应用，2021，41（4）：1055－1063.

15 杜晋华，尹浩，冯嵩．中文电子病历命名实体识别的研究与进展［J］．电子学报，2022，50（12）：3030－3053.

16 杨强．可解释人工智能导论［M］．北京：电子工业出版社，2021.

17 荣雯雯，汪刚，朱其立．基于人工智能的病历后结构化专病数据库在临床研究中的价值探讨［J］．上海交通大学学报，2020，40（7）：995－1000.

18 邵长庚．全国银屑病流行及自然病程调查［J］．医学研究通讯，1990，19（9）：29－30.

19 史潇兮，辛世杰，程帅，等．血管外科结构化电子病历设计与应用前景研究［J］．中国实用外科杂志，2021，41（3）：353－356，360.

20 吴智妍，金卫，岳路，等．电子病历命名实体识别技术研究综述［J］．计算机工程与应用，2022，58（21）：13－29.

21 万歆，姚晴虹．基于后结构化电子病历的胰腺癌科研数据平台设计［J］．医疗卫生装备，2022，43（5）：38－43，84.

22 杨慧清，胡建平，张黎黎．多中心重大出生缺陷数据共享科研大数据平台构建及应用［J］．中国卫生信息管理杂志，2023，20（4）：634－639.

23 母晓莉，徐俊，鞠伟卿，等．基于医联大数据的实时数据平台的建设与应用［J］．中国卫生信息管理杂志，2018，15（1）：70－73.

24 尚俊良，徐佳，王莒生，等．银屑病中医研究概述［J］．中医杂志，2017，58（22）：1971－1974.

25 张兰华，赵鑫，王玫，等．阶梯性区域医疗卫生资源信息一体化共享集成平台的建设与对策研究［J］．中国软科学，2022（S1）：187－192.