基于自然语言处理的肿瘤专科病历质控系统建设

刘伟伟 王立军 庞 娟 王 丹 衡反修

（北京大学肿瘤医院暨北京市肿瘤防治研究所信息技术服务部/恶性肿瘤发病机制及转化研究教育部重点实验室 北京 100142）

〔摘要〕 **目的/意义** 通过建立电子病历内涵质控系统，实现病历书写标准化与规范化，提高医院病历质量。**方法/过程** 基于医院医疗数据搭建智能中台，结合自然语言处理、机器学习技术形成具有肿瘤专科特色的知识库、规则库，实现电子病历“前置审核、全面覆盖、过程监管、闭环管理”的全新质控模式。**结果/结论** 应用基于自然语言处理的肿瘤专科病历质控系统后，质控覆盖率由 1% 提升至 100%，甲级病案率提升至 96% 以上，具有较好的实时性与准确率，为医院病历高质量发展奠定坚实的信息化基础。

〔关键词〕 内涵质控；自然语言处理；肿瘤知识库；电子病历

〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2024.02.013

Construction of a Quality Control System for Oncology Medical Records Based on Natural Language Processing

LIU Weiwei, WANG Lijun, PANG Juan, WANG Dan, HENG Fanxiu

Information Technology Service Department, Peking University Cancer Hospital & Institute/Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Beijing 100142, China

〔**Abstract**〕 **Purpose/Significance** Through the establishment of a quality control system for electronic medical record (EMR) content, the standardization and normalization of medical record writing is realized, and the quality of hospital medical record is improved. **Method/Process** The intelligent medical data center is built based on hospital medical data, and the knowledge base and rule base with tumor specialty characteristics are formed by combining natural language processing (NLP) and machine learning technology. The new quality control mode of “pre－audit, comprehensive coverage, process supervision and closed－loop management” of EMR is realized. **Result/Conclusion** After the application of the medical record quality control system based on NLP, the quality control coverage rate increased from 1% to 100%, and the rate of class A medical records increased to more than 96%, with good real－time and accuracy, providing a solid information foundation for the high－quality development of hospital medical records.

〔**Keywords**〕 content quality control; natural language processing; tumor knowledge base; electronic medical record

1 引言

电子病历系统是医疗系统与临床业务结合最紧密、临床使用最多的医疗系统之一。为提高电子病历书写质量，国家卫生监管部门相继出台《病历书

〔修回日期〕 2023－09－22

〔作者简介〕 刘伟伟，硕士，助理工程师，发表论文 1 篇；通信作者：衡反修。

写基本规范（试行）》[1]和《医疗机构病历管理规定（2013 年版）》等文件，2018 年 12 月出台的《电子病历系统应用水平分级评价管理办法（试行）》和《全国医院信息化建设标准与规范（试行）》[2]要求各地医院进一步推进病历信息化进程，提高医院医疗服务质量，对电子病历数据质量提出了更严格、具体的要求。

张坤丽等[3]应用基于规则的方法对电子病历数据进行结构化，采用最大熵模型对电子病历进行分类，以提高病历结构化的准确性，但该模型仅对首次病程记录进行去重处理及自动差异化分析，涉及病案种类较少，难以实现全覆盖。宋源等[4]基于模式层与后台数据层构建功能性胃肠病中医药知识图谱，建立较完整的胃肠病知识库，但是病历内容分词较少、知识库不够全面。马启贤[5]提出一套中文电子病历标注规则以及两种实体识别方法，提高识别与分词准确性，但是模型验证数据量有限，鲁棒性较差。

针对医院病历质量管理延迟、质控流程覆盖面窄、专科医院质控规则缺乏等问题，北京大学肿瘤医院搭建基于自然语言处理（natural language processing，NLP）技术的肿瘤专科病历质控系统，实现全院患者病历质量全流程闭环管理。该系统通过自然语言处理完成分词及语义分析，实现电子病历文书后结构化，并借助知识图谱搭建适合该院的专科类知识库、质控规则引擎库[6]，建立高效且实用的专科电子病历质控系统。

2 质控流程

2.1 传统病历质控流程

北京大学肿瘤医院电子病历系统于 2014 年上线。随后针对住院患者增补上线时限类质控功能，主要包括住院患者入院记录、日常病程时间提醒与质控，减少超时病历。终末病历依旧延用传统质控模式，见图 1。由医务部门专人抽查质控，耗时耗力；随机抽查质控容易遗漏，质控问题不全面；容易出现主观判断失误等问题。

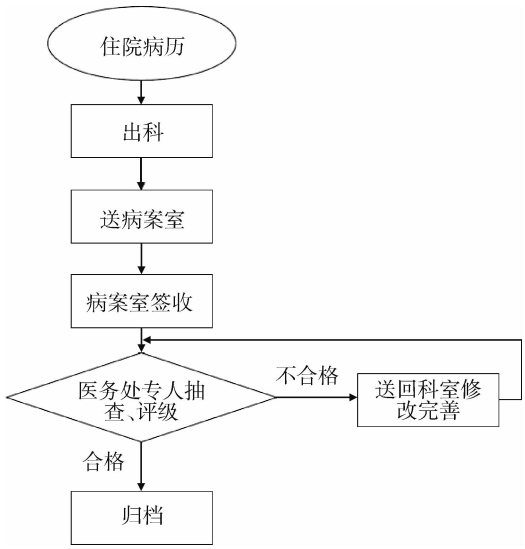

图 1 传统电子病历质控流程

2.2 人工智能电子病历质控流程（图 2）

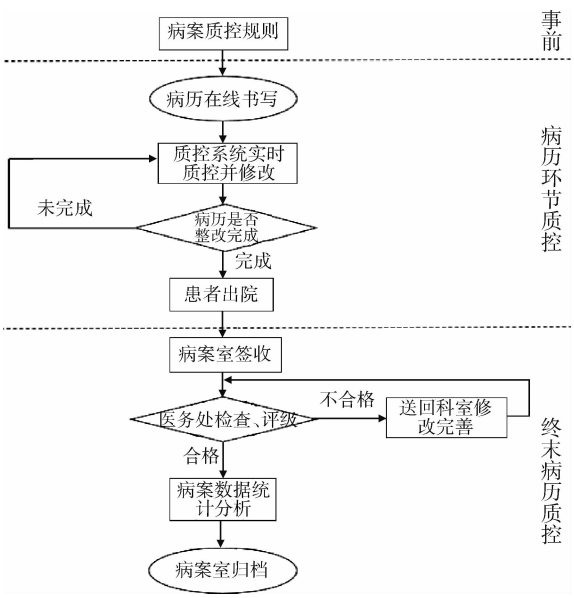

图 2 人工智能电子病历质控流程

为进一步加强医疗机构病历管理，提高病历内涵质量，助力医院高质量发展，构建以机器学习、人工智能（artificial intelligence，AI）为核心的电子病历内涵质控体系。利用自然语言处理技术，以知识库和规则库为引擎，研发“住院病历质控、门诊病历质控、病案首页质控、病案质量与核心制度监

管、肿瘤专科质控”电子病历内涵质控系统，形成电子病历“前置审核、全面覆盖、过程监管、闭环管理”的全新质控模式。实现患者病历文书全覆盖质控，实时检出病历问题并及时提醒医生修改，完成病历的前置审核与监管。患者出院且病案室签收病历后，本科室质控员与医务处质控管理员可登录质控系统针对有问题病历文书发送整改通知，医生修改后及时反馈，实现问题闭环管理可追溯。该系统可有效提高临床科室和管理部门的工作效率。

3 系统实现

电子病历内涵质控系统主要通过集成平台进行实时数据同步，辅助以抽取、转换和加载（extract – transform – load，ETL）技术抽取历史数据。数据同步后利用自然语言处理、机器学习等人工智能技术，进行分词、整合、归一等操作处理，形成大数据医院管理平台等数据应用。数据集成流程，见图 3。

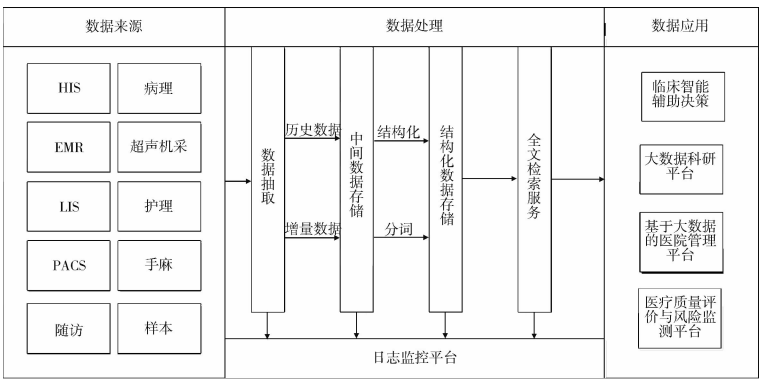

图 3 数据集成流程

3.1 数据源接入

数据来源是数据集成流程中的基础，除 HIS、检验检查等常规数据来源外，还有电子病历的异构文本数据，因此将病历文书内容后结构化处理，进行数据整合，从而形成智能医学数据中台[7]，实现各系统数据共享。每日利用数据中台同步终末病历患者的基础数据，完成质控。结合电子病历系统间实时数据交互接口进行环节病历患者数据同步，有问题及时提醒，文书每次保存后毫秒内即可完成反馈，缩短医生等候时间。

3.2 自然语言处理

面对海量的患者数据，集成流程中数据处理环节是最复杂也是最为重要的一步，运用自然语言处理技术，处理多源异构数据，实现数据的融合与汇集。结合命名实体识别与信息提取技术，识别疾病、药物等文本实体，并从中提取关键信息，如入院记录中的过敏原、肿瘤分期等。对所提取数据进行集成、清洗、分类、情感分析、规范、质量控制，从而转换成结构化可利用数据。系统归并清洗了全院近 1 年超过十万份的病历数据以及超过百万条的检验检查医嘱结构化数据。

由于医学术语主观性表达较强，同一名词在实际数据中存在不同医生采用不同书写方式的现象。利用医学知识库的数据字典将院内使用的不同书写方式进行标准化处理并保存在知识库中，在实际质控时将其映射至统一实体名称，消除书写方式不同导致的语义差异。构建符合肿瘤医院需求的肿瘤专科知识图谱，包括医疗实体、关系、属性，如疾病与症状，癌症症状包括脱发、疼痛等；药物与药品，抗肿瘤药物包括放疗药、化疗药等；疾病治疗和诊断方法，化疗、放疗以及检验、穿刺病理等。系统立足医院医疗数据，以肿瘤为核心，构建包含抗肿瘤药品、检验、放化疗等 8 类十万条规模的知识实体，11 类近百万条实体关系的知识图谱。

系统在应用自然语言技术的同时融合医院肿瘤专科特色，对部分分词切词进行医学标注，进一步

加强对肿瘤相关病历文书的分解与保存，提高准确率。以入院记录的初步诊断 TNM 分期规则分词为例，入院记录中诊断“1. 外阴皮肤恶性黑色素瘤术后 $T_{2b}N_{2a}M_{1a}$Ⅳ 期，1.1. 左腹股沟淋巴结清扫术后（3/9），1.2. 双肺多发转移；2. 周围神经病”。首先，找到诊断数据中实体并对相应实体进行标注，包括部位、程度、肿瘤、阶段、TNM 分期、临床分期；其次，根据主次诊断、肿瘤诊断分期、不同实体间的逻辑关系，进行对应语义标注与关系串联，见图 4。

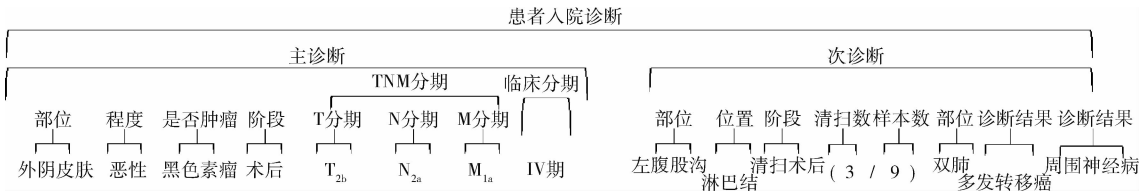

图 4 入院诊断语义分词分析

针对肿瘤专科医院特色，对入院记录、出院记录、抗肿瘤日常病程等文书，共计近千份文本数据的医学标注，制定了 14 条肿瘤专科特色质控规则，质控涉及入院记录、出院记录、日常病程、手术记录、知情同意书等各类文书。

3.3 知识库

知识库建立过程中，基于国际疾病分类（international classification of diseases，ICD）、医学系统命名法（systematized nomenclature of medicine，SNOMED）等疾病术语标准，权威指南以及医学质量管理体系、卫生信息交换标准（health level 7，HL7）等，借助临床医生帮助，建立医学专业术语与临床日常书写习惯的对应关系[8]。利用半监督机器学习方法获取初始医学数据建立本体库，借助语义之间的关联，利用机器推理和人工纠偏，参考医学相关概念和联系，形成较完整的数据层次结构，建立同一疾病间的上下层级关系表和上下语义之间的关联，其中包含概念、属性、关系和实例，以结构化形式表现。处理全院患者本体数据，形成疾病知识库、药学知识库、术语症状库、辅助检查知识库、治疗操作库、文献指南库等。全量本体数据达到 800 万条，其中院内知识库内容均达到万条级别：术语症状库超过 9 万条，疾病知识库超过 7 万条，治疗操作库超过 5 万条。

3.4 数据应用性能设计

为进一步加快数据处理速度，提高质控效率，进行如下优化。一是优化操作页面，对于响应时间要求高的操作页面，如电子病历点击保存最新数据时，均严格遵守高性能操作页面设计原则，保证使用效率。二是调整数据库，利用分布式文件存储数据库 MongoDB 的非关系型数据库，将不同类型的表存储于不同的表空间，做好不同来源数据的分类。结合用户建议和使用频率，定时将部分历史数据迁移至备份库，使每次质控的数据保持在较小数量级内。三是利用数据库连接池，作为质控系统的核心，数据库高频次打开和关闭会占用大量系统资源。利用服务器提供的数据库连接池高级特性，在系统建立之初创建若干数据库连接，使用系统时，只需快速地从连接池中得到一个已经建立好的连接即可，大大提高数据库访问速度，缩短用户等待时间。

4 应用效果

4.1 运行情况

AI 质控系统在全院运行半年以上，相较于传统质控流程，具有以下优势。一是通过优化电子病历模板减少了书写失误。针对文书必填项等要求，在制作模板时利用电子病历内嵌功能设置必选属性，“前置质控、源头治理”。重要病历模板采用结构化模式统一代码，以结构化数据集存储，提高数据质量。二是实时提醒并及时整改。在院患者采用环节质控，医生在书写保存病历时实时传输数据，可接

收小程序弹窗实时提醒，并根据提醒内容修改问题文书。科室质控员可通过质控程序查看本科室在院患者文书书写质量情况，根据问题发送整改通知，进一步提高病历质量。三是智能手段与人工结合逐步提升质控准确率。出院归档患者采用终末质控，医务处对内涵质控系统检出的乙级、丙级病历进行人工核查，减少机器误判及医疗特殊情况扣分情况，并对误判内容人工标注后再次分词处理，丰富知识库，减少误判。

4.2 运行分析

系统自上线以来，在全院各科室使用，覆盖率达 100%，远超人工质控效率，见表 1。随着质控工作的进一步深入，医务处及时收集临床使用中有关系统、质控规则、使用范围的问题，反馈并调整系统，更新机器学习模型及分词规则，不断扩大知识库，更好地适应肿瘤专科文书的质控需求。

表 1 住院病历不同质控方式比较

病历质控能力	人工质控	AI 内涵质控
病历质控数（份）	3 000	3 000
缺陷数（个/每百份）	10	100 ~ 500
病历质控覆盖率（%）	1	100
病历质控准确率（%）	99.9	93.6
质控速度（份/日）	8 ~ 15	3 000
质控时长（日）	200	1

通过环节质控实时弹窗提醒，当得分低于甲级病历时，病历扣分细节自动弹出提醒医生及时修改。2022 年 10 月弹窗功能上线以来，每份病历的点击量明显上升，见图 5，反映医生对于每份缺陷病历均能及时点击修改。由此甲级病案率由 2022 年 7 月的 88% 增长至 2022 年 12 月的 96%，见图 6，从源头解决了文书缺陷问题，大大提升了病案质量水平，减少了病案纠纷。

系统在原有规则的基础上，针对医院特色开发肿瘤专科规则，进一步满足《病案管理质量控制指标》（2021 年版）[9] 的要求，提高了医院管理效率及质量，见表 2。

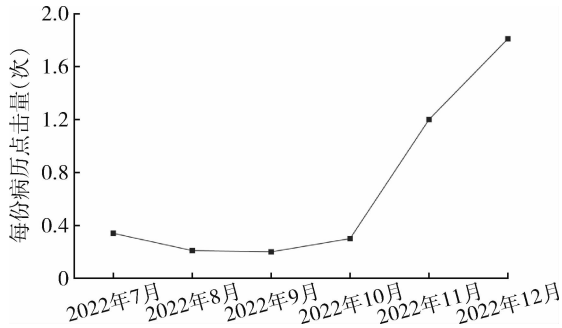

图 5 质控系统上线后每份病历点击量

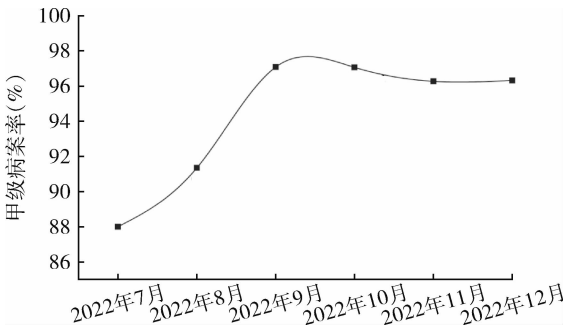

图 6 质控系统上线后甲级病案率

表 2 肿瘤专科规则上线效果

规则	2022 年 9 月（%）	2022 年 12 月（%）
首次入院患者治疗前病理诊断率	56.00	96.00
治疗前完成临床 TNM 分期率	84.00	95.00
抗肿瘤药物治疗病程规范记录率	9.20	55.00
知情同意书拟行手术方案缺陷率	4.50	1.20

5 结语

病历质控系统集成医院各业务系统医疗数据，进行数据后结构化处理，搭建医院智能数据中台，建立肿瘤专科知识库及质控规则库。搭建并推广了电子病历“前置审核、全面覆盖、过程监管、闭环管理”的质控模式，最终实现院内质控 100% 覆盖，全院甲级病案率提升至 96%，提升了全院的病案质量。

但是实际运行中尚存在病历模板结构化程度低、

（下转第 91 页）

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 国家卫生健康委员会. 全国护理事业发展规划（2021—2025 年）[EB/OL]. [2023 - 07 - 28]. https://www.gov.cn/zhengce/zhengceku/2022 - 05/09/content_5689354.htm.

2 么莉，马旭东，安磊，等. 近十年我国护理质量管理与控制工作的发展历程与展望 [J]. 中国护理管理，2022，22（12）：1761 - 1766.

3 尚文涵，李长安，吴志军，等. 国家护理质量数据平台的建设、应用及改进建议 [J]. 中国卫生质量管理，2019，26（3）：1 - 4，8.

4 高列. 青海谋划“十四五”卫生健康事业发展 [N]. 健康报，2022 - 02 - 25（3）.

5 祁少锷. 信息技术助力护理质控 [J]. 中国新技术新产品，2019（17）：34 - 35.

6 么莉. 护理敏感质量指标监测基本数据集实施指南（2018 版）[M]. 北京：人民卫生出版社，2018.

7 李茵. 面向医院管理的数据驱动决策研究 [D]. 长春：吉林大学，2021.

8 尚文涵，张海燕，么莉，等. 护理专业医疗质量控制指标（2020 年版）的构建 [J]. 中国卫生质量管理，2021，28（6）：66 - 69，74.

9 马旭东. 我国医疗质量安全不良事件分类的思考 [J]. 中国卫生质量管理，2021，28（6）：46 - 50.

10 张洁，倪平，邓欣，等. 影响出院患者满意度的关键服务指标分析 [J]. 中国卫生统计，2020，37（4）：550 - 553.

11 成守珍，高明榕，王若婧. 澳大利亚循证卫生保健中心身体约束标准介绍 [J]. 中国护理管理，2014，14（10）：1019 - 1021.

12 崔念奇，甘秀妮，张传来，等. 基于德尔菲法构建 ICU 患者身体约束评估量表 [J]. 护理学杂志，2018，33（2）：62 - 64.

13 石泽亚，杨丹，秦月兰，等. 降低 ICU 患者约束缺陷发生率的品管圈实践 [J]. 护理学报，2014，21（15）：13 - 16.

14 罗梅，张丽. 降低 ICU 患者身体约束率策略研究进展 [J]. 当代护士（上旬刊），2020，27（2）：16 - 18.

15 王晓坤，赵辉，陈永担，等. ICU 气管插管肿瘤患者下呼吸道感染的相关因素分析及护理对策 [J]. 中国医药指南，2018，16（33）：220 - 221.

16 鲁志卉，王颖，黄子菁，等. 成人气管插管非计划性拔管风险评估量表的构建 [J]. 中国护理管理，2022，22（6）：893 - 898.

（上接第 81 页）

文书内容主观表达强、后结构化分词不准确等问题。未来将进一步提高文书模板结构化覆盖率及原始数据质量，充分利用自然语言处理技术进一步加大医学数据标注量级，结合医生书写习惯及时更新知识库，从而提高分词准确率，推动医疗质量提升。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 原卫生部，国家中医药管理局. 病历书写基本规范（试行）[EB/OL]. [2023 - 01 - 28]. https://www.gov.cn/gongbao/content/2003/content_62088.htm.

2 国家卫生健康委员会. 全国医院信息化建设标准与规范（试行）[EB/OL]. [2023 - 01 - 28]. https://www.nhc.gov.cn/cms - search/xxgk/getManuscriptXxgk.htm?id = 5711872560ad4866a8f500814dcd7ddd.

3 张坤丽，马鸿超，赵悦淑，等. 基于自然语言处理的中文产科电子病历研究 [J]. 郑州大学学报（理学版），2017，49（4）：40 - 45.

4 宋源，于彤，朱玲，等. 功能性胃肠病中医药知识图谱构建方法研究 [J]. 中国数字医学，2022，17（10）：48 - 53.

5 马启贤. 中文电子病历实体识别算法研究 [D]. 兰州：西北师范大学，2021.

6 胡松林，胡雅玲. 基层医院电子病历的质控 [J]. 中国病案，2016，17（9）：20 - 22.

7 陆慧菁，杨广黔，彭俊丰，等. 基于智能医学数据中台的大数据科研平台应用实现 [J]. 中国数字医学，2020，15（4）：22 - 25.

8 朱立峰，左铭，万歆，等. 医院临床研究中的大数据应用需求与策略 [J]. 中国数字医学，2015，10（12）：14 - 15，24.

9 国家卫健委发布《病案管理质量控制指标（2021 年版）》[J]. 医学信息学杂志，2021，42（3）：94.