SHANGHAI JIAO TONG UNIVERSITY
SCHOOL OF MEDICINE

学者介绍

汪　刚 硕士
WANG Gang Master

助理研究员
Master's Degree, Assistant Professor
ORCID ID: 0000-0002-5383-2842

汪　刚（1978—），上海交通大学附属胸科医院统计中心主任。2007年获海军军医大学（原第二军医大学）临床医学硕士学位。现任上海市医院协会病案管理专业委员会委员及医院医疗保险管理专业委员会青年委员。

长期从事医院管理研究，包括结构化电子病历建设、医保智能监管、病案服务优化、临床研究专业人员培养等，探索专病数据库在医疗机构的建立、监管与使用的系列机制。主持局级课题1项。在国内核心期刊上发表论文6篇。

该研究依托上海交通大学医学院“双一流”暨高水平地方高校建设“一流学科—临床医学－临床研究中心建设”项目。

WANG Gang, born in 1978, director of the Statistical Center of Shanghai Chest Hospital, Shanghai Jiao Tong University. In 2007, he obtained his master's degree in clinical medicine from The Second Military Medical University. Currently, he is a member of Medical Record Management Professional Committee of Shanghai Hospital Association, and a member of Youth Committee of Hospital Medical Insurance Management Professional Committee of Shanghai Hospital Association.

He has long been engaged in hospital management research, including the construction of structured electronic medical records, intelligent supervision of medical insurance, optimization of medical record service, training of clinical research professionals, etc., and explored a series of mechanisms for the establishment, supervision and use of specialized disease database in medical institutions. He has presided over 1 bureau-level project and published 6 papers in domestic core journals.

The research relies on the Project of Clinical Research Center, Clinical Medicine, First-Class Discipline of "National Double First-Class" and "Shanghai-Top-Level" high education initiative at Shanghai Jiao Tong University School of Medicine.

技术与方法

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

荣雯雯[1]，汪　刚[1]，朱其立[2]
1. 上海市胸科医院，上海交通大学附属胸科医院统计中心，上海 200030；2. 上海交通大学电子信息与电气工程学院，上海 200240

[摘要] **目的·**探讨由非结构化电子病历文本信息建立的病历后结构化专病数据库在临床研究中的价值支撑。**方法·**采集 2007 年 10 月—2019 年 9 月于上海市某三甲专科医院就诊的患者信息，采用人工智能（artificial intelligence，AI）引擎等信息化方法将电子病历文本信息后结构化形成结构化数据库，并与传统结构化数据库进行对比。**结果·**采集 82 584 例患者的就诊信息，住院文书记录结构化数量 253 000 条，搭建肺癌、食管癌、纵隔肿瘤 3 个专病数据库。与传统结构化数据库相比，该专病数据库扩大了数据检索范围，提升了数据检索效率。**结论·**基于 AI 的病历后结构化专病数据库的建成，减轻了临床医师数据检索的负担，为临床研究提供了有价值的统计数据。

[关键词] 临床研究；人工智能；后结构化；数据库
[DOI] 10.3969/j.issn.1674-8115.2020.07.022　[中图分类号] R319　[文献标志码] A

Discussion on value of medical records-structured specialized disease database based on artificial intelligence in clinical research

RONG Wen-wen[1], WANG Gang[1], ZHU Qi-li[2]
1. Statistical Center, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai 200030, China; 2. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

[Abstract] Objective· To explore the value support of medical records-structured specialized disease database established by using unstructured electronic medical record text information in clinical research. **Methods·** The information of patients who were admitted to a Grade A specialist hospital in Shanghai from Oct. 2007 to Sept. 2019 were collected. By using artificial intelligence (AI) engine and other information methods, the electronic medical record text information were structured into a structured database, and compared with the traditional structured database. **Results·** The information of 82 584 patients were collected, and the structured number of hospital records was 253 000. The specialized disease databases of lung cancer, esophageal cancer and mediastinal tumor were established. Compared with the traditional structured database, the specialized disease database expanded the scope of data retrieval and improved the efficiency of data retrieval. **Conclusion·** The construction of medical records-structured specialized disease database based on AI reduces the burden of clinician data retrieval, and provides valuable statistical data for clinical research.

[Key words] clinical research; artificial intelligence (AI); structured; database

在当下的大数据时代，数据可通过挖掘来实现其自身的价值[1]。作为临床诊疗活动的重要场所，医院应当充分利用其院内海量的医疗数据，供医师开展相关临床研究，挖掘出深层次的规律[2]。有报道[3]显示，加强与重视医院的临床研究的开展，不仅可以推动临床上新技术的发展，还能够提高诊疗水平。目前，绝大多数医师收集科研数据仍需要从病案室借阅病历，再通过手工记录加以整理；即使部分医院已实现了电子病历无纸化，即将病案首页中的结构化数据整理成数据库，但对于电子病历文本中的大量非结构化数据的使用，仍需要医师通过手动来查找。一方面，手工查找费时费力，效率较低[4]；另一方面，通过该种方式使用如此海量的医疗数据，或将给临床研究造成极大的信息资源浪费。因此，如何通过人工智能（artificial intelligence，AI）实现对电子病历文本信息的有效利用，以辅助临床医师挖掘医学规律、提高临床诊疗水平成为了当下的研究热点[5]。基于此，本研究以某三甲专科医院为例，通过采用 AI 技术将电子病历文本信息结构化形成数据库，以期为临床研究的顺利开展提供价值支撑。

[基金项目] 上海交通大学科技创新专项资金（ZH2018ZDA28）。
[作者简介] 荣雯雯（1991—），女，助理统计师，硕士；电子信箱：alicedely@163.com。
[通信作者] 汪　刚，电子信箱：chestwang@163.com。
[Funding Information] Shanghai Jiao Tong University Scientific and Technological Innovation Funds (ZH2018ZDA28).
[Corresponding Author] WANG Gang, E-mail: chestwang@163.com.

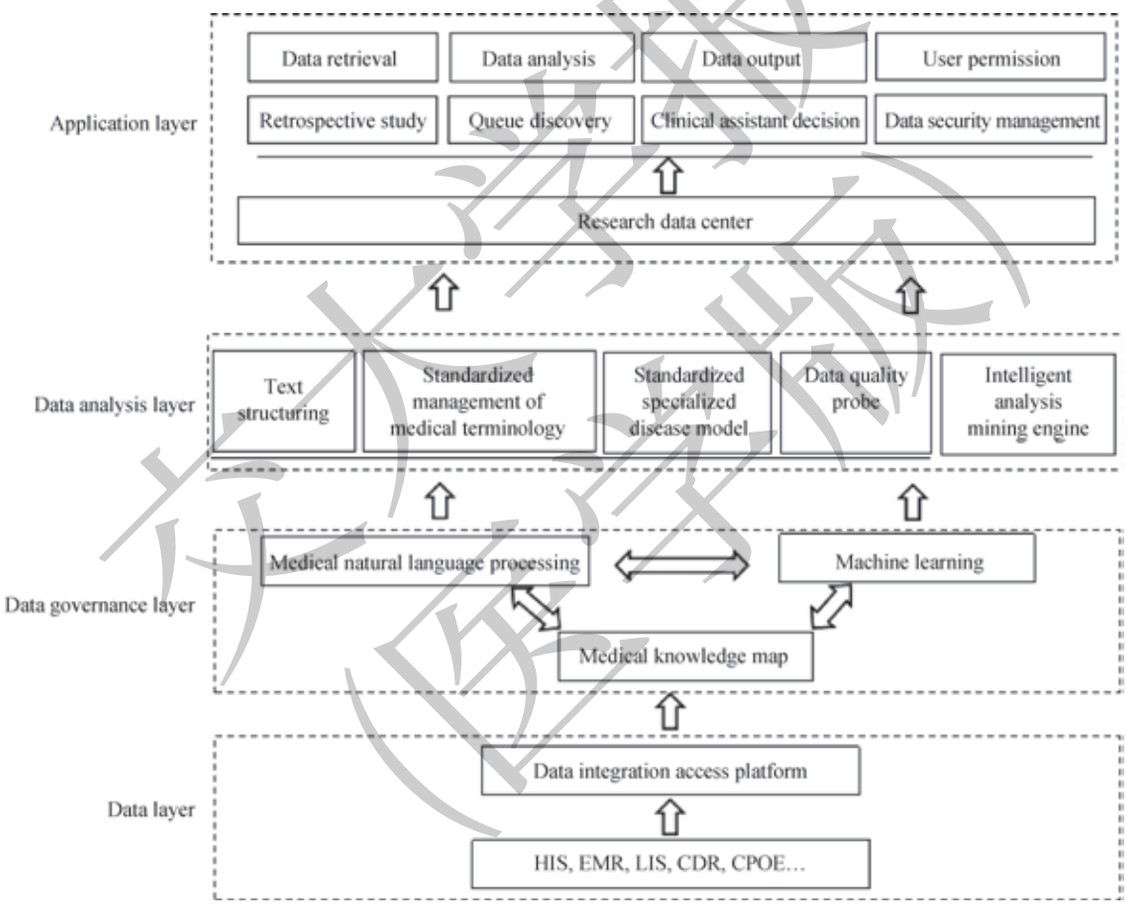
1 资料与方法

1.1 研究资料

为进一步提升某三甲专科医院临床研究的广度和深度，在保障数据安全、准确及完整的前提下，收集2007年10月—2019年9月于该三甲专科医院就诊患者的全部电子病历文本信息。

1.2 研究方法

1.2.1　专病数据库的系统设计　采用基于容器技术的分布式架构（Kubernetes，K8s）实现对专病数据库的建设。该平台能够采集医院现有的业务应用系统[如医院信息系统（hospital information system，HIS）、实验室信息系统（laboratory information system，LIS）、放射信息系统（radiology information system，RIS）、电子病历系统（electronic medical record，EMR）等]的临床数据，从而实现患者从门诊、急诊、住院及随访等的就诊、住院及预后信息的集成。其临床数据采集范围包括出院小结、病案首页、手术记录、检查检验报告、病程记录等。通过自然语言处理、知识图谱、机器学习等AI引擎实现各类医学文本数据的结构化、标准化和归一化等处理。该专病数据库的设计将为临床研究提供专病概览、智能科研检索、队列发现、科研统计分析等功能模块。其系统设计见图1。



Note: CDR—clinical data repository; CPOE—computerized physician order entry.

图1　专病数据库的系统设计流程图

Fig 1　System design flowchart of specialized disease database

1.2.2　专病数据库实现的关键技术

（1）复制技术和变更捕获技术　在专病数据库的建设过程中，需采用数据库复制技术和变更数据捕获（change data capture，CDC）技术建立实时复制库，在复制库中进行实时数据集成。数据库复制的方式包括2种，即关系型数据库SQL Server（structured query language server）利用发布订阅的方式进行复制，以及Oracle GoldenGate[6]数据复制技术。在复制数据库的同时，采用CDC技术对日志文件（任何操作都会写进其中）中发生变更的数据进行实时捕获，如增、删、改等操作。该技术会把更改应用到数据文件中，同时将符合要求的数据标记为需要添加跟踪的项。数据实时集成的技术架构见图2。

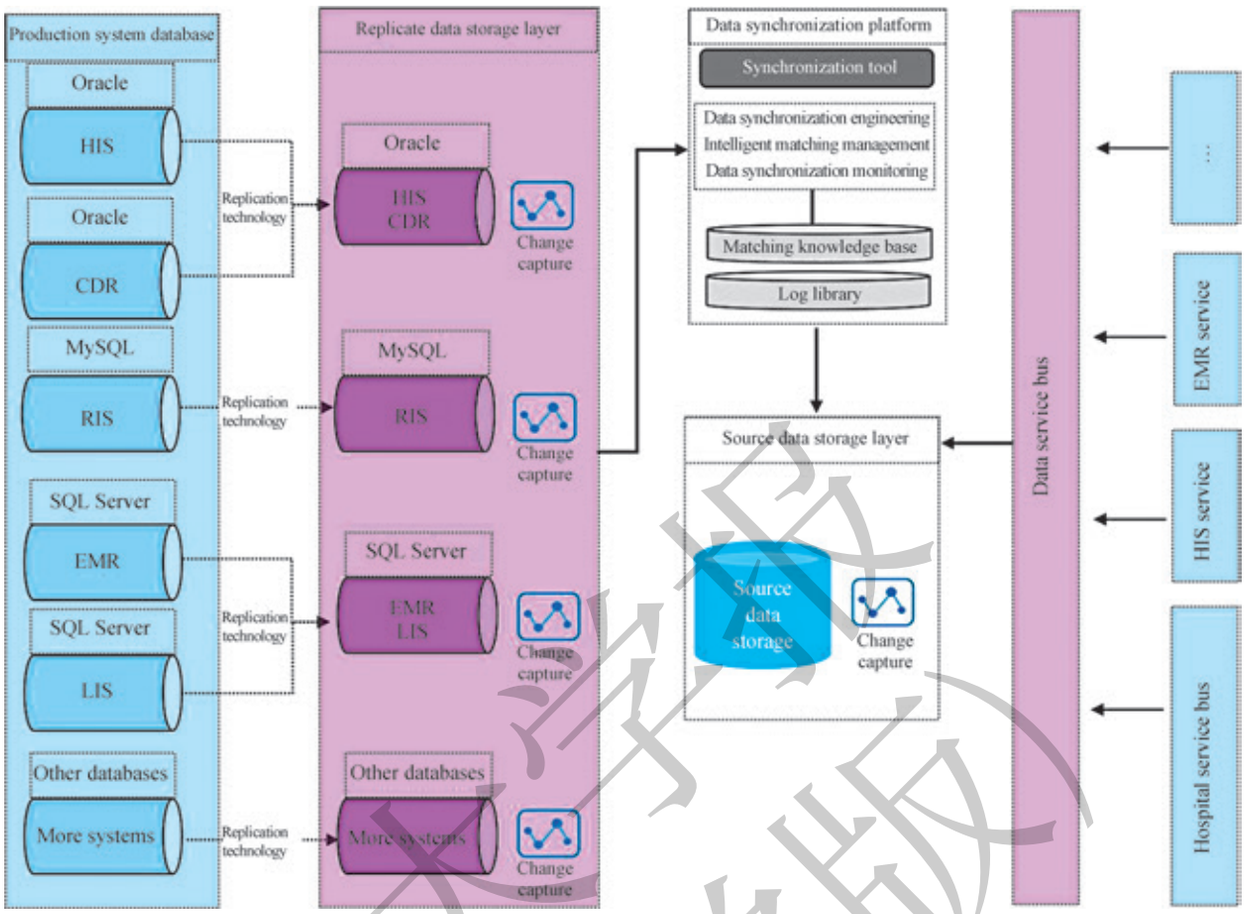


图 2 数据实时集成的技术架构图

Fig 2 Technical architecture diagram of real-time data integration

（2）医学自然语言处理　基于医院的海量病历文书，使用无监督学习、监督式学习、主动学习、迁移学习等机器学习方法建立一整套针对中文医学文本的分层式自然语言处理（natural language processing，NLP）系统，对医学文本进行信息抽取、结构化转换以及标准化处理，包括医学文本分词、医学词性标记、医学命名实体识别、实体标准化和实体关系抽取、医学文本语义依存分析等环节。①医学文本分词：对电子病历文本采用 IKAnalyzer 开源分词工具[7]，按照正向最大匹配法将文本中的字符串与充分大的机器词典的词条进行匹配。若在词典中找到某一长度的字符串，则匹配成功。②医学词性标记：采用基于规则的标注方法[8]，对电子病历文本中的每个词的词性加以标注。③医学命名实体识别：医学领域中的命名实体包括疾病名称、药物名称、检查项目名称、手术操作名称、症状、器官部位等，采用融合注意机制（Attention）的双向长短期记忆网络（bidirectional long short-term memory，Bi-LSTM）[9]设计的主动型深度学习对医学命名实体进行识别，平均精度可超过 97%。④实体标准化和实体关系抽取：采用机器学习法实现对实体标准化和实体关系的抽取。⑤医学文本语义依存分析：包括确信度分类、时序解析、关联抽取、语义树构建的整套流程，针对各种内容和类型的医学文本的行文方式建立语言学模型，并以结构学习的形式完成端对端的解析，信息抽取覆盖度占文本内包含可提取信息的 96% 以上。

（3）数据质量评估　专病数据库建成后，定期进行数据完整性和准确性评估，即根据不同病种的实际特点，采用标准化 AI 自动纠错功能，将纠错后数据与原数据进行对比查询，追溯到前端系统，以提高数据录入的准确性；同时，还需从专病数据库中随机抽调数据，与目前的病案首页系统中的数据进行比对，以确保数据的准确性。

2 结果

当前，本研究已完成肺癌、食管癌、纵隔肿瘤 3 个专病全量数据库的建设，包含 2007—2019 年肺癌就诊患者

71 263 例、食管癌就诊患者 5 883 例、纵隔肿瘤就诊患者 5 438 例，住院文书记录结构化数量 253 000 条，形成 3 个专病相关变量集，即肺癌包含 485 个变量、食管癌 559 个变量、纵隔肿瘤 481 个变量，自动填充率为 40% ～ 56%。与传统的数据库相比，该专病数据库存在如下优势：①实现了临床文本信息的后结构化，扩大了检索范围即支持全文本检索，解决了临床研究中数据采集范围受限的问题。②不仅支持按照已设定的变量进行数据检索，还支持关键字模糊检索，从而缩短了检索周期，提升了临床研究中数据检索的效率。③解决了数据沉淀不足导致无法直接使用的问题，满足了临床医生的科研需求。具体应用实例见图 3 ～图 5。

截至 2019 年底，申请使用该数据库的前 3 个科室分别为呼吸科、放疗科及肿瘤外科，申请次数分别为 9、4 和 2 次；已有多位临床医师利用专病数据库中预处理后的数据构建临床事件的预测模型，并采用机器学习的方式对疾病的发生及发展等影响因素进行多因素分析；同时，也有部分临床医师采用数据库中的数据进行临床队列研究。目前，已有临床医师利用专病数据库中的数据进行胸腔镜肺手术转开胸的危险因素及影响的研究，并成功发表文章。

图 3 专病数据库变量选择的界面
Fig 3 Interface of variable selection of specialized disease database

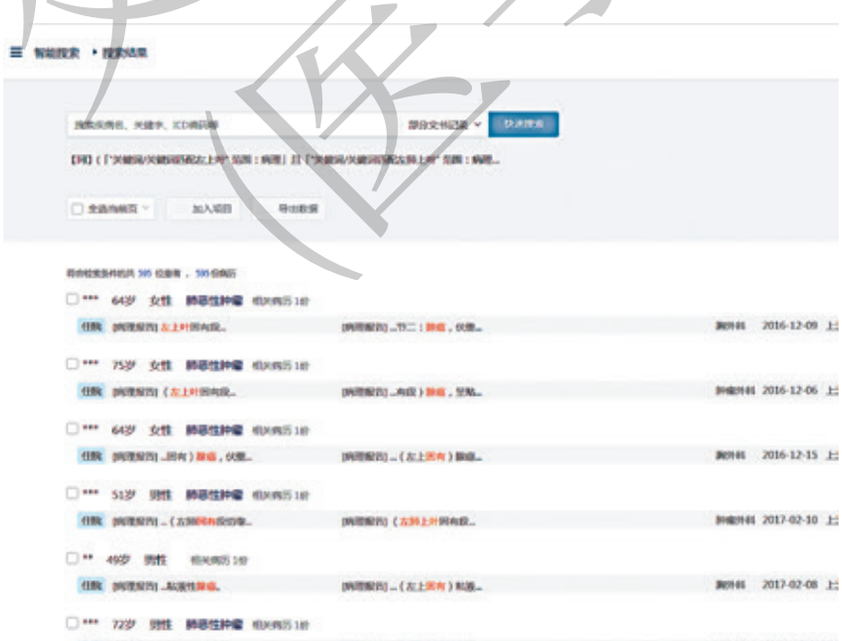


图 4 病理报告中关键词的检索结果
Fig 4 Retrieval results of key words in pathological reports

Note: PAMBA—p-aminomethylbenzoic acid.

图 5 检索结果溯源、定位显示的界面

Fig 5 Interface for tracing and displaying of search results

3 讨论

本研究就病历文本信息进行二次利用，构建专病数据库。与建设前相比，该数据库存在如下优势：①支持全文本数据检索及关键字模糊匹配检索，极大地缩短了检索周期，减轻了临床医师数据整理的负担。②检出的数据可直接用于基本的统计描述功能如性别比、年龄构成等，从而为临床研究提供了病历筛选和数据分析的模型支持，满足科研需求。③随着院外随访数据与该数据库的成功对接，可直接使用预处理后的海量原始数据进行临床队列研究训练，实现对研究对象的全面分析，获得更充分的研究结果。

然而，在专病数据库的建设过程中也遇到一些困难：①针对同一种特征描述，医师有多种写法。例如，对于阴性症状的描述，则有“否认某症状”“无某症状”“某症状(−)”“未触及某症状”等。需向 NLP 系统提供更高的提取精度、归一化术语表达，实现医学术语标准化。②提取变量时存在部分字段缺失。需通过缺失值填补形成智能化数据库，以提高数据完整性。③在建成初期，数据库系统不稳定导致数据调取时间延迟等。需及时向技术人员进行反馈并加以维护，同时需提高技术人员工作的严谨性。此外，该数据库也存在一些不足，如在数据抽取的方法上，未来可采用准确率更高的方法，即考虑结合深度学习相关的算法模型等，更加充分地利用数据本身的特征实现信息化抽取。综上，专病数据库的建设是一个不断探索的过程，需逐步积累经验、学习新的信息化技术，未来或将为临床研究提供有力的价值支撑。

参·考·文·献

[1] 刘利钊, 洪江水, 刘莉莉, 等. 面向大数据图像处理的尺度空间挖掘算法及应用 [J]. 上海交通大学学报, 2015, 49(11): 1731-1735.
[2] 王忠庆, 邵尉, 彭程, 等. 医疗大数据时代对医院统计工作的新思考 [J]. 中国卫生统计, 2015, 32(3): 542-543.
[3] 王藜箫. 加强医院科研发展与管理对提升医院核心竞争力的影响 [J]. 中国卫生产业, 2017, 14(16): 126-127.
[4] 甘霖. 基于云计算的电子病历全文检索系统 [J]. 中国数字医学, 2016, 11(12): 41-43.
[5] 彭红波, 韩晟, 王婷婷. 基于Solr的电子病历全文检索系统的设计与实现 [J]. 中国医疗设备, 2019, 34(3): 102-105.
[6] 宓正宇. 基于 Goldengate 的数据库异地灾备实现 [J]. 电信科学, 2018, 34(4): 136-143.
[7] 柴洁. 基于IKAnalyzer和Lucene的地理编码中文搜索引擎的研究与实现 [J]. 城市勘测, 2014(6): 45-50.
[8] 彭涛, 戴耀康, 朱枫彤, 等. 一种基于规则的无监督词性标注方法 [J]. 吉林大学学报 (理学版), 2015, 53(5): 956-962.
[9] 刘飞龙, 郝文宁, 陈刚, 等. 基于双线性函数注意力 Bi-LSTM 模型的机器阅读理解 [J]. 计算机科学, 2017, 44(S1): 92-96, 122.

[收稿日期] 2019-12-19 [本文编辑] 邢宇洋