电子病历信息抽取可视化分析*

娄 培 方 安

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020） （1 中国医学科学院/北京协和医学院医学信息研究所 北京 100020 2 中南大学生命科学院 长沙 410083）

赵琬清 杨晨柳 胡佳慧

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔摘要〕 以国内外电子病历信息抽取相关文献为研究样本，基于 CiteSpace 构建关键词共现图谱、时间线聚类图以及共被引网络图，可视化揭示电子病历信息抽取发展趋势、研究热点及重点领域，为相关研究提供参考。

〔关键词〕 信息抽取；电子病历；可视化；知识图谱；CiteSpace

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2021.04.007

Visualization Analysis of Electronic Medical Record Information Extraction

LOU Pei, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China; FANG An, 1Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, 2 School of Life Sciences, Central South University, Changsha 410083, China; ZHAO Wanqing, YANG Chenliu, HU Jiahui, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔**Abstract**〕 Taking relevant literatures on electronic medical record information extraction at home and abroad as study samples, the keyword co－occurrence graph, timeline clustering map and co－cited network are built based on CiteSpace to visually reveal the development trend, study hotspots and key fields of electronic medical record information extraction, so as to provide references for related study.

〔**Keywords**〕 information extraction; Electronic Medical Records (EMR); visualization; knowledge graph; CiteSpace

1 引言

临床文本中蕴含着丰富的健康医疗信息，以电子病历为代表的临床文本是医疗活动过程中产生的一种重要信息资源。电子病历以自由文本记录形式为医疗工作者撰写病历提供便利，但给临床知识发现的自动分析与获取带来巨大挑战[1]。非结构化文本数据在电子病历中占比较大，包括入院、病程、手术、出院记录等，其包含患者疾病、症状、治疗

〔修回日期〕 2020－06－10

〔作者简介〕 娄培，硕士，研究实习员；通讯作者：胡佳慧，副研究员。

〔基金项目〕 中国医学科学院中央级公益性科研院所基本科研业务费项目“面向知识发现的中文电子病历语义标注方法研究”（项目编号：2018 PT33005）；中国医学科学院医学与健康科技创新工程协同创新团队项目“中文临床医学术语系统构建研究”（项目编号：2017－I2M－3－014）。

过程等重要临床证据，亟待深入分析与挖掘。近年来全球范围内关于临床信息抽取的研究文献逐年增加，目前主流知识发现如医学知识图谱构建、疾病预测、药物预警等研究数据都来源于抽取到的结构化数据。以命名实体识别和实体关系抽取为主要研究内容的语义研究引起重视[2]。信息抽取质量将直接影响在此基础上开展的一系列临床应用深度与广度[3]。文献计量是一种定量分析方法，其利用数学和统计方法来描述和评价科学文献各种外部特征，从而预测科学技术在该领域的研究现状和发展趋势[4]。信息可视化技术使研究人员发现隐藏规则和模式变得容易，使决策更加简单[5-6]。本研究运用可视化分析工具对电子病历信息抽取领域进行文献计量和可视化分析。使用共现分析、聚类分析、中心度分析等方法生成可视化结果[7-8]，这些结果有助于直观地观察到学科发展轨迹和研究热点。

2 研究方法

为实现对电子病历信息抽取研究态势的全面分析，本文以 Web of Science 和 CNKI 数据库中收录的相关中英文文献数据为研究对象，研究框架，见图 1。

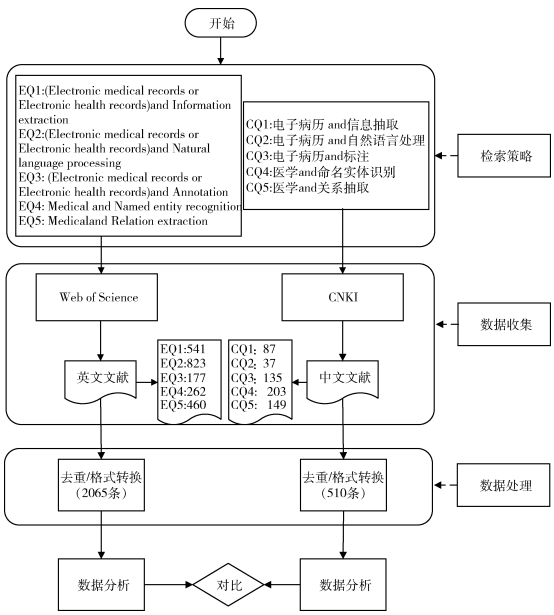

图 1 研究框架

首先根据两个不同语种数据库检索需求，分别构建 5 个中英文检索式，其中 EQ1 – EQ5 表示英文检索式，CQ1 – CQ5 表示中文检索式，设定检索年份范围为 1990 – 2020 年。根据上述策略进行文献检索，得到 2 663 条英文数据和 611 条中文数据，对检索到数据进行初步筛选，剔除与主题不相关的文献数据；在此基础上进一步对数据进行去重和格式转换，最终得到中英文文献数量分别为 510 篇和 2 065篇。在数据分析阶段，采用 CiteSpace 5.5 R2 对收集的文献数据进行可视化。

3 研究结果

3.1 总体趋势

3.1.1 概述 对中英文文献出版数量和发表时间进行跟踪，见图 2。电子病历信息抽取研究文献总体呈现增长趋势，其中英文文献在 1990 – 2010 年期间缓慢增长，2010 年后增长迅速，特别是随着 2012 年奥巴马政府正式启动大数据研发项目，带动了医疗大数据研究和应用；中文文献在 2004 – 2013 年间处于平稳发展阶段，2013 – 2016 年间随着电子病历相关管理规范[9-10]相继推行文献量开始增长，特别是在《“健康中国 2030”规划纲要》[11]《关于促进“互联网 + 医疗健康”发展的意见》[12]等一系列政策和文件推动下，应用人工智能技术开展疾病风险预测、医学影像辅助诊断、临床辅助诊疗、智能健康管理等写入医院信息化建设标准文件[13-15]，中文文献量开始迅猛增长。

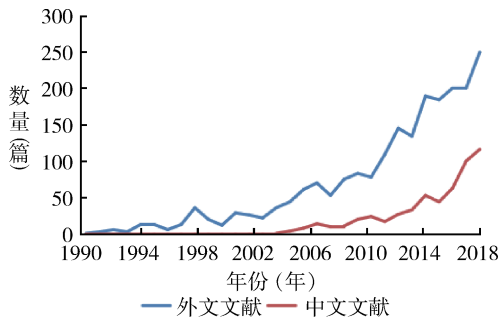

图 2 发文量统计

3.1.2 国家中心性和发文数（表 1） 美国、英国、法国、荷兰、日本中心性较高。我国电子病历发展虽然起步较晚，但在文献发表数量方面位居第 2 位，为世界范围内电子病历文本处理相关研究发

展起到积极推动作用。

表 1 国家中心性和发文数

序号	国家	中心性	文献量（篇）	序号	国家	中心性	文献量（篇）
1	美国	0.52	1 027	11	中国	0.05	186
2	英国	0.27	99	12	意大利	0.03	56
3	法国	0.26	151	13	西班牙	0.03	52
4	荷兰	0.25	46	14	罗马尼亚	0.03	20
5	日本	0.14	47	15	瑞士	0.02	50
6	德国	0.13	134	16	印度	0.02	39
7	澳大利亚	0.12	56	17	苏格兰	0.02	8
8	埃及	0.08	5	18	加拿大	0.01	65
9	新西兰	0.07	10	19	韩国	0.01	28
10	委内瑞拉	0.07	5	20	希腊	0.01	23

3.2 网络图谱

3.2.1 共现图谱 基于英文文献的关键词共现图谱，见图 3。电子病历信息抽取研究主要关注与疾病相关的医疗文本，相关任务包括文本处理、信息抽取、文本分类、文本挖掘以及系统建设等。随着机器学习等新技术涌现，将词嵌入、迁移学习、深层神经网络、表示学习等方法与临床实际需求相结合的知识发现已受到关注。基于中文文献的关键词共现图谱，见图 4。主要关注电子病历信息抽取技术，包括对中文分词、特征选择、规则构建的研究，对条件随机场、支持向量机等机器学习方法和卷积神经网络、长短期记忆网络、词向量等深度学习方法的研究。

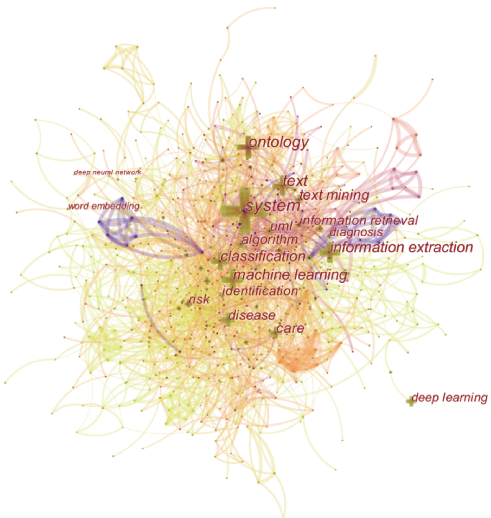

图 3 外文关键词共现网络

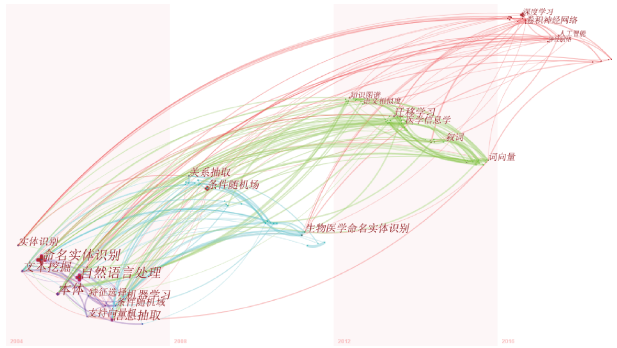

图 4 中文关键词时区

3.2.2 时间线聚类 依据时间线对关键词进行聚类，英文与中文文献聚类结果，见图 5、图 6。英文文献聚类类别包括：信息抽取类，关键词有元数据、知识表示、文本挖掘等；电子病历及临床文本类，关键词有疾病、诊断、流行病学、信息检索等；临床应用类，关键词有药物预警、药物不良反应、信息系统等；技术方法类，关键词有人工智能、机器学习、深度学习、标注等。中文文献聚类类别包括：医学本体类，关键词有本体、语义网、语义标注、语义相似度、语义分析、语义关系等；电子病历文本类，关键词有标注规范、语料库、词性标注、分词方法等；生物医学命名实体识别类，关键词包括深度学习、主动学习、迁移学习、长短期记忆网络、马尔科夫链、条件随机场等；人工智能类为近两年研究热点，主要包括知识图谱、问答系统、智慧医疗等。

3.2.3 引文网络 共被引分析是一种引文网络分析方法，通过分析共被引网络中的聚类及关键节点，可以揭示出研究领域知识结构，还可发现研究前沿、知识基础和研究演变，以及在演变过程中起到关键作用的文献[16-18]。电子病历信息抽取领域共被引可视化网络，见图 7。通过聚类得到 18 个簇，可以看出早期关注重点在语义研究上，包括语义分类、语义网等，高被引文献集中在对信息提取工具研究上。随后的高被引集群集中在自然语言处理方法，医学实体识别、深度学习、循环神经网络、卷积神经网络是这一阶段文献关注的重点。近两年主要关注利用深度学习技术对实体及关系进行抽取和大规模病历文本的机器学习技术应用，如临床决策支持系统构建、电子病历隐私保护等。

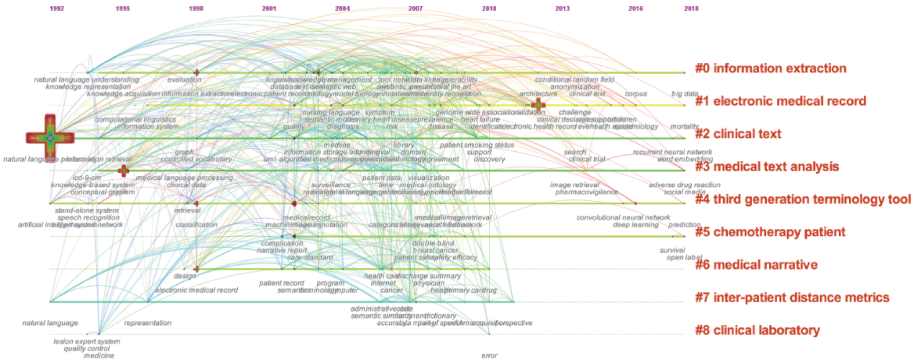

图 5 外文文献时间线聚类

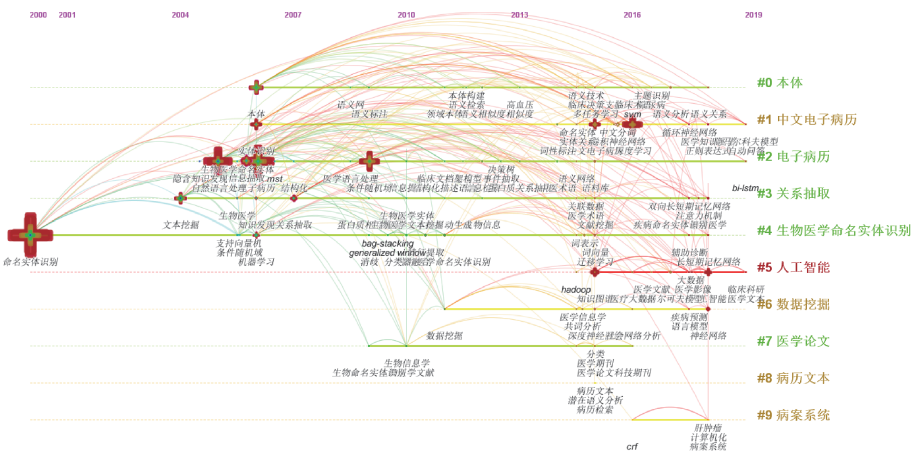

图 6 中文文献聚类

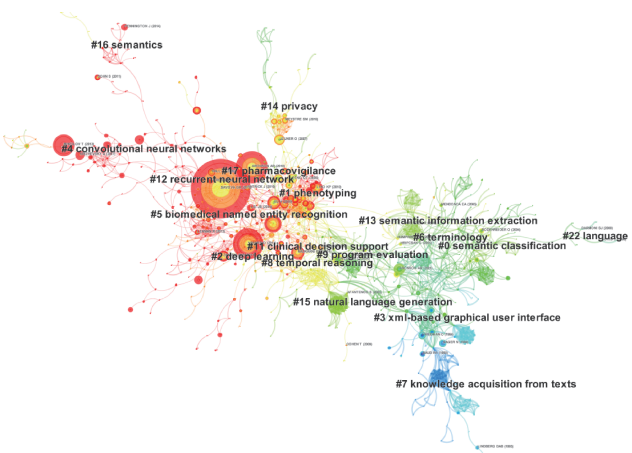

图 7 引文可视化网络

4 分析与讨论

4.1 研究热点

主要集中在自然语言处理、机器学习方法以及信息抽取模型构建。随着人工智能技术发展，深层神经网络、循环神经网络、词嵌入等方法得到关注。对比中英文文献，基于英文的电子病历相关研究起步较早，特别是美国国家集成生物与临床信息学研究中心自 2006 年起先后组织去隐私信息、概念识别及关系抽取、共指消解、特殊实体识别、风险因素识别等多个电子病历信息抽取相关任务[19]，吸引世界范围内研究者广泛参与，目前基于英文电子病历的信息抽取研究体系较为成熟，将人工智能技术应用于药物不良反应、基因疾病关系等领域的电子病历文本信息提取是研究热点。在国家相关政策推动下，基于中文电子病历的信息抽取近几年受到研究人员重视，语料库构建、临床实体及其之间关系抽取、医学知识图谱建设等相关工作已全面展开，国内知识图谱与语义计算大会于 2017 年起已连续 4 年组织电子病历信息抽取相关评测任务[20]，

中文临床命名实体及其类别和位置等属性信息提取是研究的热点，与此同时为提高信息抽取性能，机器学习算法和模型研究也是热点方向。

4.2 重点领域

4.2.1 理论方法研究 早期电子病历信息抽取任务一般采用基于统计机器学习的方法，常用模型有决策树、隐马尔科夫模型、最大熵模型、支持向量机、条件随机场等。近年来随着深度学习发展，越来越多研究将神经网络用于序列标注任务，如长短期记忆网络（Long Short Term Memory，LSTM）易于求解序列问题，已广泛用于临床实体识别任务[21]；随着注意力机制的提出，相关研究表明引入注意力机制可进一步提升模型效果[22]；此外利用迁移学习可将通用领域训练的浅层 BiLSTM 模型迁移到医学文本中，训练更深层次的神经网络来识别病历中的实体，有效提高识别性能[23]；随着 Google 在 2018 年底发布 BERT 模型，基于该模型在自然语言处理各项任务的优异表现，越来越多的研究将该模型用于电子病历信息抽取[24]。

4.2.2 系统工具研发 面向临床信息处理需要，信息抽取系统及工具研发也是重点。临床文本自然语言处理系统（Clinical Text Analysis and Knowledge Extraction System，cTAKES）是最具代表性的模块化、可扩展开源自然语言处理系统，使用梅奥诊所的标注数据，利用规则和机器学习结合的模型进行实体识别、本体映射、主题识别等[25]。药品提取系统（Medication Extraction System，MedEx）使用出院小结记录进行训练，可对临床记录中的药物名称进行识别[26]，系统在药品名称标准化、药物不良事件挖掘等方面都有广泛应用。健康信息文本提取工具（Health Information Text Extraction，HITEx）可对呼吸系统疾病名称、并发症和吸烟状况进行提取[27]。

4.2.3 临床应用研究 将先进算法、模型与临床实际应用需求相结合是研究重点，例如疾病预测、个性化诊疗、药物预警、医疗流程优化等。在疾病预测方面，利用卷积神经网络和长短时记忆网络训练败血症早期发现模型，可在早期提示患病风险并促进干预[28]。利用循环神经网络对电子病历数据进行表征学习，提取数据时序特征，可对心血管疾病进行风险预测[29]。利用 Logistic 回归选取特征变量，可构建基于支持向量机的重度急性胰腺炎早期预警模型[30]。在个性化诊疗方面，基于注意力机制的长短时记忆网络训练病历语料发现其内在特征，可为不同疾病患者提供个性化药物治疗方案[31]。对电子病历进行信息提取，构建知识图谱，利用知识图谱图结构关系设计问答系统，可用于临床辅助诊疗[32]。

5 结语

本文以国内外电子病历信息抽取相关文献为研究样本，基于 CiteSpace 构建关键词共现图谱、时间线聚类图以及共被引网络，对其发展趋势、研究热点及重点领域进行分析。研究表明电子病历信息抽取是临床文本数据深度处理与应用的基础，也是临床语料库和医学知识图谱等的重点研究内容，包括命名实体识别、实体属性及关系抽取，需要综合利用自然语言处理、机器学习等先进算法、模型及相关工具提升信息抽取质量和效率。中文电子病历信息抽取目前尚处于起步阶段，具有极大研究潜力和广阔应用前景。面向我国临床应用实际需求，可进一步加强中文临床文本处理与信息抽取理论方法研究与探索；充分利用人工智能等先进技术，不断提升研究深度与广度；注重科技成果转化，逐步提升我国电子病历信息抽取国际化水平。

参考文献

1 胡佳慧，方安，赵琬清，等. 面向知识发现的中文电子病历标注方法研究［J］. 数据分析与知识发现，2019，3（7）：123－132.

2 王建洪. 中文电子病历信息提取方法研究［D］. 长沙：湖南大学，2016.

3 张晗，郭渊博，李涛. 结合 GAN 与 BiLSTM－Attention－CRF 的领域命名实体识别［J］. 计算机研究与发展，2019，56（9）：1851－1858.

4 方志蓉. 学术期刊的文献计量指标及提升其数值的途径探求［J］. 出版发行研究，2005（4）：66－68.

5 Yuran J，Xin L. Visualizing the Hotspots and Emerging Trends

of Multimedia Big Data through Scientometrics [J]. Multimedia Tools and Applications, 2019 (78): 1289 - 1313.

6 董献洲, 胡晓峰, 司光亚. 信息可视化技术在情报分析中的应用研究 [J]. 计算机工程与应用, 2006, 42 (34): 175 - 177.

7 Yibing Chen, Xiaofang Tong, Junge Ren, et al. Current Research Trends in Traditional Chinese Medicine Formula: a bibliometric review from 2000 to 2016 [EB/OL]. [2019 - 03 - 03]. https://pubmed.ncbi.nlm.nih.gov/30941195/.

8 Miao Y, Liu R, Pu Y, et al. Trends in Esophageal and Esophagogastric Junction Cancer Research from 2007 to 2016 [J]. Medicine, 2017, 96 (20): e6924.

9 原卫生部. 电子病历基本规范 (试行) [EB/OL]. [2010 - 02 - 22]. http://www.gov.cn/zwgk/2010 - 03/04/content_1547432.htm.

10 杨睿, 李婧, 马兆辉, 等. 中医电子病历基本数据集标准的研究思路与方法 [J]. 中国数字医学, 2017, 12 (5): 74 - 76.

11 《中国肿瘤》编辑部. "健康中国 2030" 规划纲要 [J]. 中国肿瘤, 2019, 28 (10): 724.

12 王玉霞, 路杰, 姚进文, 等. "互联网 + 医疗健康" 平台建设与应用实践 [J]. 医学信息学杂志, 2019, 40 (3): 25 - 30.

13 《医学信息学杂志》编辑部. 国家卫健委答复: 人工智能辅助诊疗平台如何提升基层医疗服务能力 [J]. 医学信息学杂志, 2019, 40 (1): 93.

14 李华才. 依托标准建设医院信息化的战略抉择 [J]. 中国数字医学, 2018, 13 (5): 1.

15 《中国卫生信息管理杂志》编辑部. 中国卫生信息技术/健康医疗大数据应用交流大会暨软硬件与健康医疗产品展览会在西安隆重开幕 [J]. 中国卫生信息管理杂志, 2019, 16 (4): 382.

16 Miao Y, Zhang Y, Yin L. Trends in Hepatocellular Carcinoma Research from 2008 to 2017: a bibliometric analysis [J]. Peer J, 2018 (6): e5477.

17 Shengqi Chen, Ruixue Bie, Yunfeng Lai, et al. Trends and Development in Enteral Nutrition Application for Ventilator Associated Pneumonia: a scientometric research study (1996 - 2018) [EB/OL]. [2020 - 02 - 03]. https://www.zhangqiaokeyan.com/academic - journal - foreign - pmc_detail_thesis/040005197209.html.

18 Gu D, Li J, Li X, et al. Visualizing the Knowledge Structure and Evolution of Big Data Research in Healthcare Informatics [J]. International Journal of Medical Informatics, 2017 (98): 22 - 32.

19 Cossin Sébastien, Lebrun Luc, Aymeric Niamkey, et al. SmartCRF: a prototype to visualize, search and annotate an electronic health record from an i2b2 clinical data warehouse [J]. Studies in Health Technology and Informatics, 2019 (264): 1445 - 1446.

20 孙安, 于英香, 罗永刚, 等. 序列标注模型中的字粒度特征提取方案研究——以 CCKS2017: Task2 临床病历命名实体识别任务为例 [J]. 图书情报工作, 2018, 62 (11): 103 - 111.

21 Habibi M, Weber L, Neves M, et al. Deep Learning with Word Embeddings Improves Biomedical Named Entity Recognition [J]. Bioinformatics, 2017, 33 (14): i37 - i48.

22 於张闲, 胡孔法. 基于 BERT - Att - biLSTM 模型的医学信息分类研究 [J]. 计算机时代, 2020 (3): 1 - 4.

23 Dong X. Formation and Containment Control for High - Order Linear Swarm Systems [M]. Berlin: Springer Publishing Company Incorporated, 2015.

24 Xiangyang Li, Huan Zhang, Xiaohua Zhou. Chinese Clinical Named Entity Recognition with Variant Neural Structures Based on BERT Methods [J]. Journal of Biomedical Informatics, 2020 (107): 103422.

25 Savova G K, Masanz J J, Ogren P V, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications [J]. Journal of the American Medical Informatics Association, 2010, 17 (5): 507 - 513.

26 Xu H, Stenner S P, Doan S, et al. MedEx: a medication information extraction system for clinical narratives [J]. Journal of the American Medical Informatics Association, 2010, 17 (1): 19 - 24.

27 Fiszman M, Chapman W W, Aronsky D, et al. Automatic Detection of Acute Bacterial Pneumonia from Chest X - ray Reports [J]. Journal of the American Medical Informatics Association, 2000, 7 (6): 593 - 604.

28 Lauritsen S M, Kalør M E, Kongsgaard E L, et al. Early Detection of Sepsis Utilizing Deep Learning on Electronic Health Record Event Sequences [EB/OL]. [2019 - 06 - 10]. https://www.researchgate.net/publication/333674698_Early_detection_of_sepsis_utilizing_deep_learning_on_electronic_health_record_event_sequences.

29 安莹, 黄能军, 杨荣, 等. 基于深度学习的心血管疾病风险预测模型 [J]. 中国医学物理学杂志, 2019, 36 (9): 1103 - 1112.

30 张晔, 张晗, 尹玢璨, 等. 基于电子病历利用支持向量机构建疾病预测模型——以重度急性胰腺炎早期预警为例 [J]. 现代图书情报技术, 2016 (2): 83 - 89.

31 梁治钢, 王一敏. 深度学习在电子病历抗菌药物使用方法分类中的应用 [J]. 计算机系统应用, 2019, 28 (8): 71 - 77.

32 杨笑然. 基于知识图谱的医疗专家系统 [D]. 杭州: 浙江大学, 2018.