基于知识图谱的 MIMIC 电子病历数据集的研究热点和趋势分析

刘 萍[1],张嘉鹏[2]

(1.中山大学附属第七医院重症医学科,广东 深圳 518000;
2.嘉应学院医学院,广东 梅州 514021)

摘要:目的 了解 MIMIC 电子病历数据集国内外研究热点和趋势。**方法** 利用 CiteSpace、HistCite 和 VOSviewer 软件,从发文量、发文机构、主要研究者、高被引期刊、高被引论文、研究主题 6 个纬度,对 CBKI、Web of Science 中有关 MIMIC 电子病历数据集的研究文献进行文献计量分析。**结果** MIMIC 领域内的发文量逐年递增;主要发文机构是美国的麻省理工学院和中国的浙江大学;主要研究作者是美国学者 CELI LA;同行内高被引期刊为 *CRITICAL CARE MEDICINE*;高被引论文是编号 37 号文献;研究主题 2018 年前主要是以 MIMIC 数据集患者行为研究为主,2019 年以后主要集中于患者死亡危险因素分析;关键词聚类分析前期研究主要与患者心肺肾系统疾病相关,后期集中于急性肾损伤和 MIMIC 数据库数据完善研究。**结论** MIMIC 电子病历数据集的研究机构和研究者之间合作较少,国内相关研究也较少;研究主题局限于 ICU 患者死亡风险的预测和生存率的分析,应扩展主题研究;未来 MIMIC 研究趋势将侧重于结合 AI 技术深度挖掘临床电子病历数据。

关键词:MIMIC;知识图谱;CiteSpace;HistCite;VOSviewer;电子病历

中图分类号:R319 **文献标识码:**A **DOI:**10.3969/j.issn.1006-1959.2023.07.012

文章编号:1006-1959(2023)07-0067-07

Research Hotspots and Trend Analysis of MIMIC Electronic Medical Record Data Set Based on Knowledge Graph

LIU Ping[1],ZHANG Jia-peng[2]

(1.Department of Critical Medicine,the Seventh Affiliated Hospital,Sun Yat-sen University,Shenzhen 518000,Guangdong,China;
2.College of Medcine,Jiaying University,Meizhou 514021,Guangdong,China)

Abstract: Objective To understand the research hotspots and trends of MIMIC electronic medical record data sets at home and abroad. **Methods** CiteSpace, HistCite and VosViewer software were used to conduct a bibliometric analysis of the research literature on the MIMIC electronic medical record data set in CBKI and Web of Science from six latitudes: number of publications, institutions, major researchers, highly cited journals, highly cited papers, and research topics.**Results** The number of publications in the field of MIMIC was increasing year by year; the main publishing institutions were Massachusetts Institute of Technology in the United States and Zhejiang University in China; the main research author was American scholar CELI LA; the highly cited journal in the peer was *CRITICAL CARE MEDICINE*; highly cited paper was No.37 literature; the research topic was mainly based on the patient behavior research of MIMIC data set before 2018, and mainly focused on the analysis of patient death risk factors after 2019. Keyword cluster analysis showed that the previous studies were mainly related to cardiopulmonary and renal system diseases, and later studies were focused on acute kidney injury and MIMIC database data improvement.**Conclusion** There is less cooperation between research institutions and researchers in the MIMIC electronic medical record data set, and there are few related studies in China. The research topic is limited to the prediction of death risk and survival rate of ICU patients, and the research should be expanded. The future research trend of MIMIC will focus on the deep mining of clinical electronic medical record data with AI technology.

Key words:MIMIC;Knowledge map;CiteSpace;HistCite;VOSviewer;Electronic health record

随着医疗技术和大数据的发展,传统纸质版病历已经不能满足医院需求,电子病历系统应运而生。电子病历数据是日常诊疗过程中数字化存档的医学数据,也是医院核心信息系统[1],主要包括患者的个人信息、病程记录、检验结果、医嘱、手术记录、护理记录、影像学检查等信息。电子病历主要优势是内容充分、书写标准规范、检索使用便利、存储简易,能辅助临床诊断,对远程医疗更有帮助,可提供快速、便捷、准确的患者资料[2]。同时,电子病历的利用可以产生巨大的临床效益, 临床随机对照试验(randomized controlled trials,RCT)是可靠性最高的“金标准”,但其耗费巨大的人力物力和财力,经过漫长的

作者简介:刘萍(1994.8-),女,湖南衡东县人,本科,技师,主要从事重症呼吸研究

通讯作者:张嘉鹏(1992.6-),男,广东五华县人,硕士,讲师,主要从事急危重症研究

周期后,研究结果却可能无法适用于所有的患者人群[3]。对于门诊和住院病房中出现的实际临床问题,即使是严格控制的 RCT,也难以捕捉到细节完整、相互影响和关联的各种信息。随着大数据挖掘技术的发展,科研工作者对医疗大数据的挖掘关注度和参与度越来越高,但是我国电子病历起步较晚,缺乏对国内外医疗大数据内容的梳理以及对研究热点、发展趋势的分析[4]。美国大型电子病历数据集 MIMIC 建立于 2003 年,是麻省理工学院、飞利浦医疗以及贝斯以色列女执事医学中心共同合作的一个生物工程项目,该项目获得了美国国立生物医学影像和生物工程研究所的资助[5]。MIMIC 数据库包含了贝斯以色列女执事医学中心(美国顶级的学术型医疗中心)所有的内外科 ICU 患者数据。目前发布的是第四代数据库(MIMIC-IV),包含了超过 4 万例患者的数据,有数千个变量。MIMIC-IV 数据库对患者信息进行了去标识化处理,并进行了标注,向研究人员免费开放共享。除了医院患者的基本信息,MIMIC-IV 数据库还包含详细的生理和临床数据。该数据库可应用于重症监护领域的大数据研究,还可用于开发和评估先进的 ICU 患者监护和决策支持系统提高重症监护领域的临床决策的效率、准确性和时效性[6]。因此,本研究以美国大型电子病历数据集 MIMIC 为例,深入剖析其研究内容,利用文献计量分析软件进行分析,以发现该数据集的相关研究在内容和广度、研究主题分布、研究热点趋势、领域内主要研究作者、机构和国家等方面的问题,以期为国内科研者深入开展电子病历的研究提供方向和指导。

1 资料与方法

1.1 数据来源 中文数据来源于中国学术期刊网络出版总库(CNKI)。检索式为:(主题=MIMIC 数据)OR(主题=重症监护医学数据库),检索时限为 2004年 1 月至 2021 年 6 月。外文数据来源于美国科学情报研究所研发的 Web of Science TM 核心合集数据库,检索式为:(主题“MIMIC-II”)OR(主题“Medical Information Mart for Intensive Care”)OR(主题“MIMIC-III”)OR(主题“MIMIC-IV”),检索时限为 2004-2022 年。

1.2 纳入与排除标准 纳入标准:研究主题关于 MIMIC 原始研究或综述文献。排除标准:报纸、会议论文等非研究型文献;无作者文献;检索主题词不符合的文献。

1.3 文献筛选与资料提取 将收集的文献导出格式为纯文本格式(.txt),通过 NoteExpress 软件依照文献纳入和排除标准筛选文献,提取文题、作者、发表期刊、发表年份、关键词等题录信息。

1.4 统计学方法 将纳入文献导入 CiteSpace 软件进行数据分析,在参数设置中,文献研究时长跨度设置为 2004~2021 年,按照分析需求,将时间切片设置为 2 年,节点类型分别选择作者(author)、机构(institution)、国家(country)、关键词(key word)、共被引期刊(cited journal)等,阈值选择 top20%,网络裁剪功能区选择参数 Pathfinder、Pruning sliced networks 或者 Puring the merge network 进行图谱分析。

2 结果

2.1 发文量分析 学术论文的年度发文量在一定程度上可以反映该研究领域所处阶段、现状和发展趋势[7,8]。LCS(local citation scores)是指某一文献在本地数据集中的被引用次数,由于 HistCite 导入的文献都是和检索词有关系的,因此,这些文章可以认为是该文献领域内的研究同行。如果某一篇文献的 LCS 值很高,就意味着它是该研究领域内的重要文献,或者是开创性文献[9]。而 GCS(globle citation scores)是某一文献在 Web of Science 核心合集中总共被引用的次数[10]。由于 2021 年的数据还没有最终补充更新完成,所以从发文量看 2021 年比 2020 年略有下降。对 2004-2021 年国家发文情况进行分析,发现 Top5 发文量的国家有美国、中国、英国、印度、加拿大,见表 1。其中,美国发文量排第 1 位,共有 282 篇,LCS 次数为 543 次,GCS 次数为 5072 次,两项指标均名列前茅;中国发文量排第 2 位,为 243 篇,同行引用次数为 110 次,总被引用次数 955 次。2004-2021 年关于 MIMIC 领域发文量见图 1,通过统计分析发现发文量总体呈现上升趋势,说明越来越受到学者重视。

表 1 Top5 国家年总发文量

排名	国家	发文量	LCS	GSC
1	美国	282	543	5072
2	中国	243	110	955
3	英国	48	215	1180
4	印度	46	16	151
5	加拿大	34	16	223

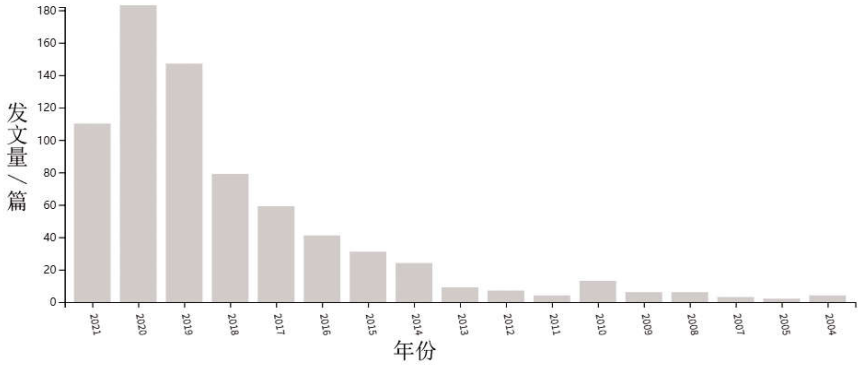


图 1 年度发文量

2.2 作者分析 通常在评价科学家的科技成果时，与其在重要刊物上发表论文的数量高度相关[11]。因此，发表论文的数量可以作为评价科学人才的一种依据[12]。该领域共有 139 名学者，发文量排在前 5 的作者见表 2，按照发文量排名，美国学者 CELI LA 排在第 1 名。作者合作图谱中总体呈现“一超多强”的合作关系，作者之间合作主要以同地区或同单位合作为主，而跨机构跨地区之间的交流甚少，见图 2。

表 2 Top5 高产作者和机构

排名	作者	来源机构	年份
1	CELI LA	美国-麻省理工学院	2016
2	ZHANG ZH	中国-浙江大学	2014
3	SAEED M	美国-密歇根大学	2004
4	LI Y	美国-欧道明大学	2020
5	YU Y	马来西亚-拉曼大学	2019

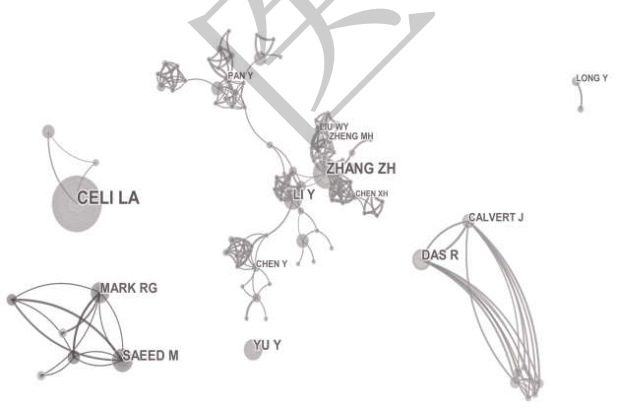


图 2 主要作者合作关系图谱

2.3 研究机构分析 发文量排在前 5 的研究机构如表 2 所示，不同机构之间的学术合作能推动该学科的发展[13]。美国麻省理工学院是所有研究机构中发文量排名第一的机构，在关系图中也是作为“中心枢纽”，连接着其他研究机构，例如哈佛以医学院、贝斯以色列女执事医学中心，但与其他不同国家之间的合作缺乏；国内以浙江大学的 ZHANG ZH 为该领域内的主要研究机构和学者，其中与国内温州医科大学、清华大学、中山大学的合作紧密，见图 3。通过节点的连线数量和大小来判读不同组群的强度及其相互作用关系，可以发现跨国之间的合作缺乏，究其原因除地理差距外，语言障碍和资金项目是主要障碍。

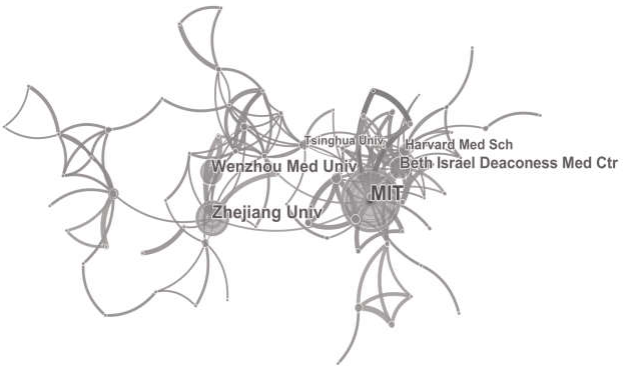


图 3 主要机构合作关系图谱

2.4 期刊分布分析 该领域共有 345 家期刊参与，根据 LCS 排名，同行被引率前四位的期刊是：*CRIT CARE MED*；*JAM MED INFORM ASSN*；*J BIOMED INFORM*；*COMPUTERS IN BIOLOGY AND MEDICINE*，见表 3。发文量和合作关系最紧密的最多的期刊分别是 *SCI DATA* 和 CRIT，见图 4。圆圈面积越大表示发文量越多，圆圈之间的连续越多表示期刊之间联系越紧密。

表 3 TLCS 排名前 4 期刊

排序	期刊	发文量	LCS	GSC
1	*CRITICAL CARE MEDICINE*	8	149	709
2	*JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*	18	63	301
3	*JOURNAL OF BIOMEDICAL INFORMATICS*	23	53	386
4	*COMPUTERS IN BIOLOGY AND MEDICINE*	10	27	157

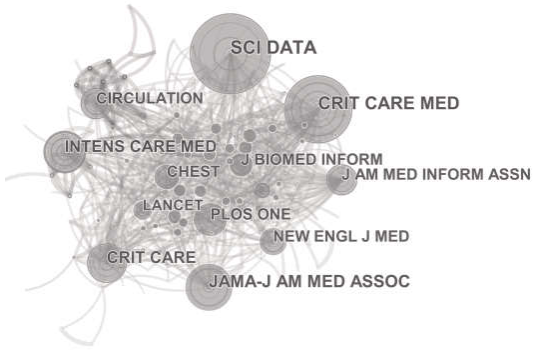


图 4 主要期刊关系图谱

2.5 被引情况分析 高被引论文是一个领域内重要的知识来源，它反映了该领域研究的重点、发展方向和研究前沿[9]。利用 Histcite 软件对 LCS 排名前 20 的论文进行共被引关系展示，圆圈越大表示文献被引频次越高，箭头表示文献之间的引证关系。本研究依次对排名前 5 的文献进行展开分析，其中编号 37、240 的文献影响力最大，编号 93、137 的文献影响力次之，编号 258 的文献影响力紧随其后，见图 5。

2.6 研究热点分析 利用 VOSviewer 和 CiteSpace 软件进行主题词分析，构建 MIMIC 领域的主题知识图谱，通过主题词频率出现的多少来反映该领域研究的热点和未来趋势。依据标题和摘要提取出的主题词，出现频率不小于 20 次，共 122 个主题词符合条件，热点分布见图 6A。图 6B 连线数量代表主题词之间共现频率，频率越高说明该主题词热度越高，从左至右代表主题词热点出现时间远近。由图可见，2018 年以前主要研究热点集中于 MIMC 领域内的行为研究（performance），2019 年以后主要研究热点集中于死亡率（mortality）的研究。

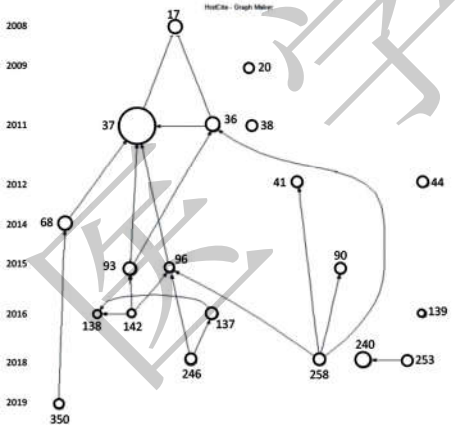
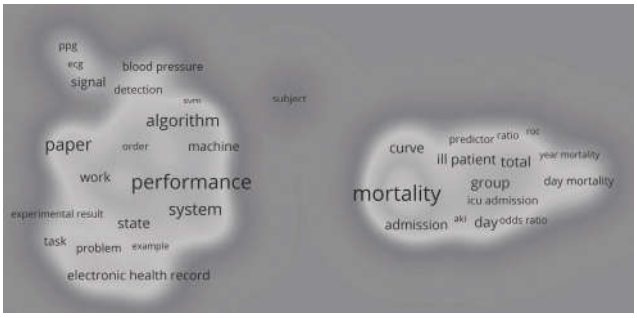
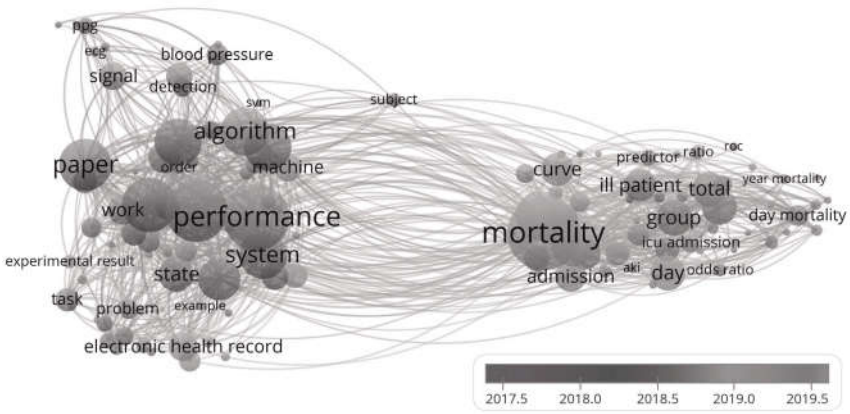


图 5 MIMIC 领域相关文献高被引论文之间的引文关系图谱



A

图 6 MIMIC 领域主题次密度图谱



B

图 6 MIMIC 领域主题次密度图谱(续)

2.7 关键词分析 对关键词聚类分析后，共生成 14大聚类，按照时间顺序，分为前中后 3 个阶段，前期聚类有 #11aki~#13intensive；中期聚类集中于 #9all-cause mortality、#6medical services、#7central venous pressure、#4albumin；后期聚类集中于 #2acute kidney injury、#1data models、#3mimic iii database，见图 7。对关键词突现分析，去除检索词后，突现强度排名前 5 的分别为：system、intensive care、clinical decision support system、medical informatics，见图 8。

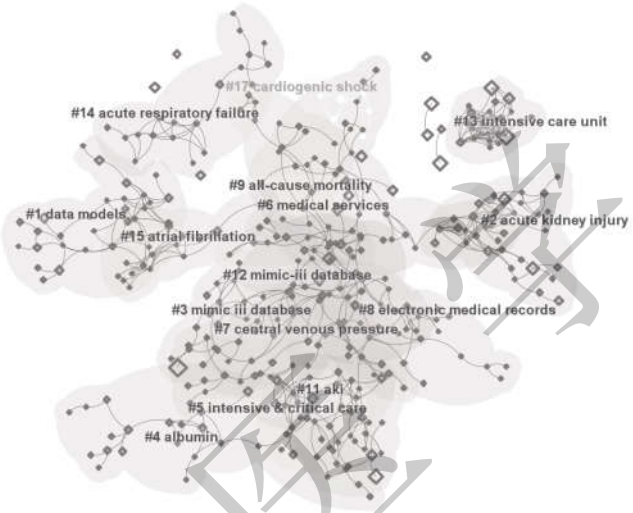


图 7 MIMIC 关键词聚类

Top 16 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2004 - 2022
system	2008	5.69	2008	2019	
mimic ii	2008	3.87	2008	2017	
decision support	2008	3.72	2008	2019	
intensive care	2013	5.24	2013	2019	
feature selection	2012	3.5	2012	2017	
heart rate	2015	3.16	2015	2021	
clinical decision support system	2016	3.87	2016	2017	
medical informatics	2016	3.83	2016	2017	
severe sepsis	2016	3.3	2016	2019	
extraction	2017	3.22	2017	2021	
definition	2016	3.22	2016	2019	
signal	2016	3.13	2016	2019	
acute physiology	2015	3.1	2016	2019	
mortality prediction	2016	3.09	2016	2019	
natural language processing	2018	3.61	2018	2021	
electronic health record	2014	3.14	2018	2021	

图 8 MIMIC 关键词突现

3 讨论

MIMIC 是一个开放的数据库，世界各地研究人员经过适当的培训以及通过相关资质测试后，就可以使用该数据库，目前已经发布了第四代数据库，包含了内外科 ICU 的患者数据，包含了超过 4 万例患者的数据，有数千个数据变量[5]。从本研究结果来看，目前研究 MIMIC 数据库的国家较多，主要集中于美国和中国，虽然美国发文量仅比中国多 39 篇，但其 LCS(543)和 GCS(5072)明显高于排名第 2 的中国，说明重要的研究成果主要来自于美国。

其中又以高校科研机构为主要力量，例如美国的麻省理工学院、贝斯以色列女执事医疗中心和哈佛大学等，由于 MIMIC 数据库是由麻省理工学院和贝斯以色列女执事医疗中心以及飞利浦医疗共同发布，因此，主要研究机构和科研人员都集中于此。而国内主要以浙江大学、温州医学院和中山大学为主要研究机构，相同地区跨机构合作较频繁，但跨国家间的合作比较少，主要研究者也是来自高校或科研机构。

从 LCS 高被引文献结果来看，编号 37[14]和编号

240[15]的文献均是论述 MIMIC 数据库信息组成和代码应用，可见，临床电子数据的信息集成管理和利用分析是当前医疗大数据应用的研究重点，主要研究热点集中于 MIMIC 数据库的推广应用。这与主题词的分布热点相符合。我国医疗行业电子病历数据库的应用起步较晚，国外目前比较成熟的电子病历数据库除 MIMIC 外，还有：美国建立的以患者为中心的结局研究网络（National Patient-Centered clinical research network，PCORnet）；英格兰地区国家医疗服务机构（National Health Services，NHS），该数据库是世界上最大人口健康数据库之一，存储了英格兰地区所有人口终生的医疗信息记录；Eicu-Philip 数据库，该数据库是由飞利浦与麻省理工学院合作推出的；美国退伍军人健康信息系统（Veterans Health Information Systems and Technology Architecture，VIstA）数据库是运用于美国退伍军人事务部的医疗系统；国家手术品质改善计划（National Surgical Quality Improvement Program，NSQUIP）是美国外科医师学会牵头的一个国际项目，该数据库收集大量外科、患者预后及不良事件相关数据。而国内类似于 MIMIC 的数据库尚未成熟，仍处于发展阶段。

文献共被引次数最多的前 5 篇论文，主要研究内容如下：编号 37 文献[14]系统介绍了 MIMIC-II 数据库包含的内容；编号 240 文献[15]主要概述了 MIMIC 数据的常用代码，主要代码包括疾病严重程度的评分表，并发症、败血症的治疗和病理生理检查、器官衰竭评分表以及治疗方案等；编号 93 文献[16]探讨了 SICULA 评价系统相比较于传统的评价系统（SAPS-II、APACHE-II、SOFA），对 ICU 脓毒血症患者死亡风险预测的效果。编号 137 的文献[17]探讨了 InSight 评价系统用于对 ICU 脓毒血症患者的评价效果，最终发现 InSight 系统在预测评价 ICU 脓毒血症患者的效果优于其他 5 种传统评价方法，且该系统的稳健性良好。编号 258 文献[18]通过纳入 MIMIC-Ⅲ数据库的患者，结果表明了深度神经网络模型在预测评估 ICU 患者临床风险率方面始终优于传统预测模型，尤其是当患者人数较多，以入院时间作为其中暴露因素的条件下，深度神经网络模型具有较高的基准性和稳健性。

MIMIC 研究主题比较集中，主要集中于两大类，一是 MIMIC 数据库的行为研究，二是 ICU 患者死亡率的预测及其危险因素的分析。随着人工智能技术的发展，结合 AI 技术的专科大数据应用和深度神经网络模型的发展，预计未来国内关注的方向主要集中于利用 AI 技术和机器模型与深度学习技术，在拥有丰富医疗大数据的 MIMIC 数据库中提取数据集构建知识库，可以更好地为临床决策者提供参考；或者针对 ICU 常见疾病做危险因素分析[19]，有利于医护人员更好地指导患者避免危险因素，这也与编号 258 文献[18]的研究结果相一致。除此之外，关于 MIMIC 领域内的其他研究主题较少，建议在后续研究中，拓展其他方向的研究。

关键词是一篇论文研究主题的高度概括，对于高频关键词的分析可以反映某一个研究领域的热点主题[20]。根据关键词聚类分析和突现分析可见，MIMIC 数据的研究热点早期主要集中于心肺重症患者的研究；到后期研究热点集中于数据的数据库系统的完善、急性肾衰竭的研究。MIMIC 数据库在建立之初，由于纳入患者数据不完善，研究多以心肺系统重症患者为主，后来随着 MIMIC-Ⅲ和 MIMIC-Ⅳ数据库的建立，越来越多的研究病例纳入该数据库中，因此，后期研究热点范围变广，包括急性肾损伤、生化常规指标（白蛋白、中心静脉压等）。

本研究存在的不足之处：由于 CitesPace 和 VOSviewer 软件无法同时对多个数据库（例如 PubMed 和 Web of Science）进行分析，而 HistCite 软件只能针对 Web of Science 核心数据库分析，因此，可能导致纳入文献有所遗漏。希望后续研究者能在此基础上纳入更多数据库的文献进行分析。

参考文献：

[1]陈静，李保萍.MIMIC-Ⅲ电子病历数据集及其挖掘研究[J].信息资源管理学报，2017(4):29-37.

[2]杨荣根，王博，龚乐君.基于 CRF 和深度学习的病历实体识别的研究[J].南京师范大学学报（工程技术版），2022,22(1):81-85.

[3]李杰.胸部术后不同区域镇痛技术的镇痛效果和安全性：随机对照试验的网络荟萃分析 [D]. 乌鲁木齐: 新疆医科大学, 2022.

[4]王国睿.基于文本挖掘的电子病历研究现状分析及热点发现[D].太原：山西医科大学，2022.

[5]Scott DJ,Lee J,Silva I,et al.Accessing the public MIMIC-II intensive care relational database for clinical research [J].BMC Med Inform Decis Mak,2013,13:9.

[6]Lee J,Scott DJ,Villarroel M,et al.Open-access MIMIC-II database for intensive care research [J].Annu Int Conf IEEE Eng

Med Biol Soc,2011,2011:8315-8318.
[7]刘祥茂,戚曼.基于社会网络分析的我国马拉松热点问题及发展趋势研究[J].四川体育科学,2023,42(1):36-42.
[8]欧敏,邹霞.自适应学习研究现状及趋势分析[J].西部素质教育,2023,9(2):23-27.
[9]祝清松,冷伏海.基于引文内容分析的高被引论文主题识别研究[J].中国图书馆学报,2014(1):39-49.
[10]田军.信息可视化分析工具的比较分析——以 CiteSpace、HistCite 和 RefViz 为例[J].图书馆学研究,2014(14):90-95,54.
[11]庞龙.科学引文分析的科学评价功能和意义[D].太原:山西大学,2006.
[12]Sloan DA,Donnelly MB,Schwartz RW,et al.The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance [J].Ann Surg,1995,222(6):735-742.
[13]刘国兵,常芳玲.基于 CiteSpace 的国内语料库翻译学研究知识图谱分析[J].河南师范大学学报(自然科学版),2018,46(6):111-120.
[14]Saeed M,Villarroel M,Reisner AT,et al.Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database[J].Crit Care Med,2011,39(5):952-960.
[15]Johnson AE,Stone DJ,Celi LA,et al.The MIMIC Code Repository: enabling reproducibility in critical care research[J].J Am Med Inform Assoc,2018,25(1):32-39.
[16]Pirracchio R,Petersen ML,Carone M,et al.Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study [J].Lancet Respir Med,2015,3(1):42-52.
[17]Desautels T,Calvert J,Hoffman J,et al.Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach[J].JMIR Med Inform,2016,4(3):e28.
[18]Purushotham S,Meng C,Che Z,et al.Benchmarking deep learning models on large healthcare datasets[J].J Biomed Inform,2018,83:112-134.
[19]Fuchs L,Chronaki CE,Park S,et al.ICU admission characteristics and mortality rates among elderly and very elderly patients [J].Intensive Care Med,2012,38(10):1654-1661.
[20]王晓晓,郭清.基于 CiteSpace 的近十年我国医养结合研究热点及发展趋势分析[J].中国全科医学,2021,24(1):92-97.

收稿日期:2023-02-08;修回日期:2023-03-03
编辑/王萌