MIMIC-III 电子病历数据集及其挖掘研究

陈　静[1]　李保萍[2]
(1.华中师范大学信息管理学院,武汉,430079;　2.武汉大学信息管理学院,武汉,430072)

[摘　要]　为了解美国典型开源医疗数据库－重症监护室电子病例数据集(MIMIC)内容及其研究利用情况,本文系统调查梳理 MIMIC 数据集的内容,从发文数量、发文国家机构、研究主题及研究方法四个维度,对 Web of Science 中的有关研究文献进行文献计量与内容剖析。结果表明,对 MIMIC 数据集的研究利用处于上升时期,但研究深度与广度不足;我国高校与医疗公司应加强对其研究利用;研究主题侧重 ICU 病人预后与死亡率预测,应扩展研究主题;挖掘方法偏好通用机器学习和统计分析方法,应加强针对性挖掘方法研究。

[关键词]　MIMIC　电子病历　数据挖掘　主题分布　重症加强护理病房(ICU)

[中图分类号]　G352.1　[文献标识码]　A　[文章编号]　2095-2171(2017)04-0029-09
DOI:10.13365/j.jirm.2017.04.029

Research on MIMIC-III Electronic Medical Record Dataset and Its Mining

Chen Jing[1]　Li Baoping[2]
(1.School of Information Management, Central China Normal University, Wuhan, 430079;
2.School of Information Management, Wuhan University, Wuhan, 430072)

[Abstract]　In order to understand the content and the research and utilization of a typical and open source medical database in the United States－the Medical Information Mart for Intensive Care, also called MIMIC, this paper systematically investigates the content of the MIMIC dataset and analyzes the literatures about the research of the MIMIC dataset in the Web of Science, respectively from number of documents, nations, institutions, research topics and research methods. It is found that the research and utilization of MIMIC dataset are in the ascending period, and its research depth and breadth are insufficient. Universities and medical companies in China should strengthen their researches and utilization on MIMIC dataset. The research topics focus on the prediction of prognosis and mortality of ICU patients, and researchers may explore more research topics. Data mining methods mainly focus on the general machine learning and statistical analysis methods, while mining methods with stronger pertinence need to be proposed.

[Key words]　MIMIC　Electronic medical record　Data mining　Topic distribution　Intensive care unit (ICU)

[基金项目]　本文系教育部人文社会科学重点研究基地重大项目“大数据资源的挖掘与服务研究——面向医疗健康领域”、湖北省高校省级教学研究项目“信息管理类‘知识主题－课程’体系网络构建研究”(2016078)的成果之一。
[作者简介]　陈静,女,副教授,研究方向为信息组织与信息检索;李保萍(通讯作者),女,研究生,研究方向为信息分析,Email:1640351499@qq.com。

1 引言

电子病历系统是医院核心信息系统[1]，随着卫生信息化建设的推进和大数据的发展，电子病历系统中积累的医疗大数据迅速增长。同时，对医疗健康大数据的有效利用可以产生巨大的经济效益，据麦肯锡预测，医疗健康数据的有效利用每年可为美国医疗健康体系带来3000多亿美元的潜在价值，贡献0.7%的年度生产力增长；可为加拿大医疗健康体系节省100亿美元的卫生费用，相当于加拿大2012年卫生总费用的5%[2]。随着大数据技术的快速发展，国际科研工作者对医疗大数据挖掘研究的关注度和参与度越来越高，但是，相关研究还缺乏针对医疗大数据内容的梳理与深入剖析，不利于科学认识医疗大数据挖掘研究现状与存在问题并促进其深化发展。因此，本文以典型医院医疗大数据－重症监护室电子病例数据集MIMIC为例，深度剖析医院医疗大数据内容，对其研究利用情况进行文献计量分析，以发现该数据集研究在内容深度与广度、参研国家机构、研究主题及研究方法等方面的不足，为深化发展电子病历大数据挖掘研究提供方向指导。

2 MIMIC-Ⅲ数据库的获取

多参数智能监测数据库(MIMIC-III)是一个免费开放的、公共资源的重症监护室研究数据库。该数据库于2006年由美国麻省理工学院计算生理学实验室以及贝斯以色列迪康医学中心(BIDMC)和飞利浦医疗共同发布，吸引了越来越多的学术界和工业界的研究人员采用该医疗数据库从事医疗研究。

MIMIC数据集包括MIMIC-II数据集和MIMIC-III数据集，MIMIC-II数据集的数据是2001—2008年间贝斯以色列迪康医学中心(BIDMC)重症监护室中病人的医疗数据，MIMIC-III数据集的数据是2001年6月—2012年10月重症监护室病人数据。数据集MIMIC数据库从发布到现在，随着更多数据变得可用，数据导入和提取方法的改进，以及数据库维护人员一直根据社区提供的数据库内容的反馈定期更新数据集，因此MIMIC数据集有多个版本，目前最新的版本是2016年9月发布的MIMIC-III V1.4。

为了方便研究人员更容易查看获取数据库，麻省理工学院计算生理学实验室提供了两个主要的软件工具：基于Web的在线访问工具QueryBuilder和可下载的虚拟机(VM)映像，两种工具都是免费开放提供给合法用户的[3]。QueryBuilder为使用者提供数据库的概况信息，研究者通过电脑端或移动Web浏览器访问数据库，利用SQL语句快速检索数据库中的信息，探索数据库中各种表和视图的结构，并检查它们之间的关系，确定所得信息是否满足研究需要，但是为了防止用户过度消耗QueryBuilder上的共享资源(例如导出MIMIC中的所有表)，导致服务器过载，该访问方式限制提供检索结果的前5000行，若检索结果超过15分钟，则检索失败。然而越来越多的用户希望运行更复杂的查询，已经开始导致QueryBuilder超载。为了缓解这个问题，麻省理工学院计算生理学实验室提供了可下载的虚拟机(VM)，允许用户在自己的计算机上运行MIMIC关系数据库的副本，VM是完全隔离的操作系统安装，可以在主机环境中运行。数据库副本包含了该数据库的所有信息，用户可以直接访问本机查找需要的所有数据，便于用户更快更方便地访问数据库。

MIMIC数据库为关系数据库，支持SQL语言查询，可以将数据集导入大型关系型数据库如SQL Server、Postgres、MySQL、Oracle进行数据处理。同时数据集中的表数据均可以以.CSV的方式导出，可利用Excel、Spss、Matlab等大型统计软件进行数据处理和统计分析辅助研究。

3 MIMIC-Ⅲ数据库内容

MIMIC数据集，早期名为多参数重症智能监测系统数据集(the Multi-parameter Intelligent Monitoring for Intensive Care)，现在名为监护室医学信息数据集(the Medical Information Mart for Intensive Care)，是一个基于重症监护室病人监测情况的医学开源数据集。其公布的目的在于促进医学研究，提升ICU决策支持水平。

从2001—2012年间，MIMIC数据集共获

得了贝斯以色列迪康医学中心重症监护室中超过50000位病人的医疗信息(如生命体征、化验结果、用药情况等)、生物图像(如超声波图像、核磁共振检测图像、CT图像等)、医疗过程及人口统计信息(入院出院时间、年龄、身高、是否死亡等)。现在数据的最新版本为MIMIC-Ⅲ,数据集主要包含两类三个数据库,分别是医院数据库和ICU数据库,医院数据库包含病人在医院里的个人信息和相关治疗信息,ICU数据库包括CareVue ICU数据库和Metavision ICU数据库,CareVue ICU数据库包含的是2001到2008年病人在ICU里由CareVue监测系统获得的关于病人治疗的相关信息,Metavision ICU数据库包含的是2008—2012年病人在ICU里由Metavision监测系统获得的关于病人治疗的相关信息。

MIMIC数据库中包含了多种类型ICU(外科监护室、内科监护室、创伤外科监护室、新生儿监护室、心脏病监护室、心外恢复监护室)。MIMIC-Ⅲ数据集主要包括波形数据集(病人的生命体征趋势图)和临床数据集,按照记录内容的不同,共包含以下21个数据表:住院表、出院表、当前使用医疗服务记录表(CPT)、日期型事件表、医务人员表、监测情况表、ICD病情确诊表、诊断相关组编码表(DRG)、ICU记录表、注射记录表(CV)、注射记录表(MV)、排泄记录表、化验记录表、微生物检测记录表、文本报告记录表、病人登记表、处方信息表、过程事件表(MV)、ICD手术记录表、服务表、病房转移表。同时,数据集中还包含了5个辅助表用来辅助查找:目前使用医疗服务术语表、ICD病情确诊词典表、ICD医疗过程词典表、ICU化验词典表、门诊化验词典表。

在对26个数据表的内容充分了解后,按照各个表的内容相关程度可分为四类,分别是病人基本信息及转移信息表、病人医院门诊的治疗相关信息表、病人在ICU里的治疗相关信息表和辅助信息表。下面分别介绍数据表的主要内容和利用该数据表进行的相关研究。

表1 病人基本信息及转移信息表

数据表名称	内容	相关研究
PATIENTS (病人登记表)	关于病人的基本信息,包含病人的性别、出生日期以及死亡日期	大部分研究都会用到此表
ADMISSIONS (住院表)	包含病人入院、出院以及死亡时间,以及人口统计信息,种族、语言、宗教、婚姻状态等	大部分研究都会用到此表
CALLOUT (出院表)	提供病人准备从ICU出院或已经出院的相关信息,如出院前病房、出院后的病房、以及出院结果等	2014, Dejam A等[4] 2013,Mayaud L等[5]
ICUSTAYS (ICU记录表)	关于病人进出ICU的记录信息,包含ICU的类型、病房号、进出ICU的时间以及时间长短	2014, Niemi M等[6] 2014,Fuchs L等[7]
TRANSFERS (病房转移表)	病人在医院期间病房转移表,主要包括转移前后的病房、进出时间、转移状态、住院时长	2014, Niemi M等[6]
SERVICES (服务表)	病人在医院期间接受的治疗,包括之前的治疗、现在的治疗以及转换的时间	2014,Moskowitz A等[8]

表1梳理了病人基本信息及转移信息表,包括病人登记表、住院、出院表以及病房转移表等,大部分关于MIMIC数据集的研究都会用到病人登记表和住院表,研究人员利用这类信息表研究影响ICU病人死亡率的因素以及预测死亡率,如Dejam等人主要研究年龄、临床背景对危重病人红细胞输血的结果和死亡率的影响,结果表明红细胞输血对整体研究人群的死亡率没有影响,但是对不同年龄段的死亡率有影响,红细胞输血对老年人有益,对年轻患者无益[4]。

表2梳理了病人医院门诊治疗的相关信息表,包括病人的诊断信息、测量的各项指标信息以及在门诊接受的医疗服务信息和医生为病人开的处方信息表,研究人员利用这类信息表主要进行优化药物用量和某些测量因素

和疾病之间是否有相关关系的研究。Ghassemi 等人以肝素药物施用为例，利用逻辑双因素回归分析提出了一种优化药物用量的方法[12]；Moskowitz A 等人利用回顾性队列研究方法探究病人入住 ICU 前镁浓度和乳酸性酸中毒的关系，研究表明低镁血症与乳酸性酸中毒的风险增加相关[8]。

表 3 梳理了病人在 ICU 里的治疗相关信息表，包括医务人员信息表、化验记录表、日期型事件表、注射事件表和医疗过程事件表，研究人员主要利用这类信息表研究病人的体征信息和 ICU 死亡率之间的关系或者与某种疾病发病率的关系，如 Sabinai 等人利用 Logistic 多变量回归分析的方法探究红细胞宽度和危重病人 ICU 死亡率之间的关系，证明了红细胞宽度是 ICU 预后重要指标，提高了 ICU 的风险

表 2　病人医院门诊治疗的相关信息表

数据表名称	内容	相关研究
CPTEVENTS（当前使用医疗服务记录表）	关于病人在医院获得 CPT 的记录信息，包括 CPT 编码、记录时间等	大部分研究都会用到此表
DIAGNOSES_ICD（诊断信息表）	根据 ICD_9 标准的病人确诊信息，包含病人编号、ICD_9 编码	2009，Goldstein I 等人[9] 2014，Lee J 等人[6]
DRGCODES（诊断相关组编码表）	病人的诊断信息类型信息表，包括 DRG 编码，DRG 类型信息	大部分研究都会用到此表
LABEVENTS（门诊检查记录表）	病人在门诊科室测量的项目记录，包括项目 ID、测量值、测量时间	2011，Celi LAG 等人[10] 2012，Kothari R 等人[11] 2014，Moskowitz A 等人[8]
MICROBIOLOGYEVENTS（微生物检测记录表）	检测病人对微生物是否过敏的信息记录表，包括测量样本 ID、类型、描述以及测量时间等	
PRESCRIPTIONS（处方信息表）	处方医生为病人开的处方用药表，包括处方有效的截止时间、药物类型、药物名称、药量	2014，Ghassemi MM 等人[12]

表 3　病人在 ICU 里的治疗相关信息表

数据表名称	内容	相关研究
CAREGIVERS（医务人员信息表）	关于医务人员的信息表，包含医务人员编号和类型	
CHARTEVENTS（化验记录表）	病人在 ICU 期间体征测量信息，包含测量项目、测量人员、测量值、测量单位、测量时间等	2012，Sabina Hunziker 等人[13] 2012，Saeed M 等人[14]
DATETIMEEVENTS（日期型事件表）	病人在 ICU 期间所有测量项目的日期，包含测量项目、测量人员、测量值、测量单位、测量时间等	2012，Lehman LH 等人[15] 2012，Hug C 等人[16]
INPUTEVENTS_CV（注射事件表(CV)）	病人在 ICU 期间由 CV 系统监测的药物注射情况，包括医务人员编号、注射量、注射速率、注射开始和结束时间	
INPUTEVENTS_MV（注射事件表(MV)）	病人在 ICU 期间由 MV 系统监测的药物注射情况，包括医务人员编号、注射量、注射速率、注射开始和结束时间	
NOTEEVENTS（文本记录事件表）	提供病人相关记录信息，包括护理记录、影像报告和出院记录等	2008，Ishna Neamatullah 等人[17] 2014，Lee J 等人[5]
OUTPUTEVENTS（排泄记录表）	病人在 ICU 期间排泄记录，包括项目名称、排泄量、排泄时间等信息	
PROCEDUREEVENTS_MV（医疗过程事件表）	病人在 ICU 内由 MV 系统监测的治疗记录表，包含项目名称、开始结束时间、项目测量值以及医务人员编号等信息	
PROCEDURES_ICD（ICD 手术记录表）	ICU 全部已完成手术的粗略信息，主要包含手术 ICD_9 编码	2008，Ishna Neamatullah 等人[15] 2012，Saeed M 等人[14]

预测 SAPS 评分[13]。Saeed 等进行多变量逻辑回归发现低血压和急性肾损伤(AKI)之间的相关性,研究结果表明,AKI 的风险与低血压的严重程度相关[14]。

表 4 梳理了相关的辅助信息表,辅助信息表都是数据词典表,用于辅助解释其他信息表,ICU 化验词典表和门诊化验词典表记录病人在 ICU 和门诊测量各项指标的详细信息,目前使用的医疗服务术语表、ICD 病情确诊词典表和 ICD 医疗过程词典表三个辅助表,是维护人员根据国家医疗标准整理得到的数据词典表,方便研究人员使用。

表 4 辅助信息表

数据表名称	内容	相关研究
D_CPT (目前使用医疗服务术语表)	主要介绍医疗服务术语,医疗服务主要分为 8 个类别,包括:评估和管理、麻醉、外科、放射科、病理和实验室、内科、新兴技术、药品以及测量	大部分研究都会用到此表
D_ICD_DIAGNOSES (ICD 病情确诊词典表)	主要包括医生为病人诊断的疾病简称以及全称	大部分研究都会用到此表
D_ICD_PROCEDURES (ICD 医疗过程词典表)	主要包括病人接收手术治疗的简称以及全称	大部分研究都会用到此表
D_ITEMS (ICU 化验词典表)	主要包括病人在 ICU 中化验项目 ID、名称、缩写、来源以及类型信息	大部分研究都会用到此表
D_LABITEMS (门诊化验词典表)	主要包括病人在门诊科室化验项目 ID 以及类型信息	大部分研究都会用到此表

4 基于 MIMIC 数据集的相关研究

MIMIC 数据集发布以来,就受到学术界和工业界研究人员的欢迎,研究人员利用 MIMIC 数据集做了很多研究,为了总结这些研究的现状、热点及趋势等,本文通过对利用 MIMIC 数据集进行研究的国内外文献进行分析,以图表的形式展示该领域研究的相关情况。

4.1 数据来源与研究方法

由于关于 MIMIC 数据集研究的中文文献较少,因此本文以 Web of Science 核心集作为文献来源,在 Web of Science 核心集中以主题为检索项,检索词为 MIMIC-II、MIMIC-III、Multi-parameter Intelligent Monitoring for Intensive Care,连接词为 or,时间跨度为 2004 年—至今,语种为 English,其他项均为默认值,共得到 122 篇文献。由于文献量较少,采用人工提取相关信息,统计论文的发文量、发文国家及研究结构、从论文摘要模块提取研究主题、从论文数据处理方法模块提取研究方法,辅以 Excel 进行数据整理分析,然后从论文的每年发文量、发文国家及研究机构、研究主题以及研究方法 4 个维度分析关于 MIMIC 数据集的相关研究文献。

4.2 关于 MIMIC 数据集研究的基础分析

4.2.1 发文量及时间分布

一个研究领域发展可以通过各年份的论文数量来体现。图 1 中折线代表 Web of Science 核心刊关于 MIMIC 数据集研究的发文量,我们可以看出,对 MIMIC 数据集的文献最早出现于 2004 年,2016 年发文量最多,为 27 篇,2004—2006 年发文量较多,2006—2010 年发文量总体呈现较快增长趋势,2011—2016 年发文量呈现逐年急剧上升趋势,从总体趋势来看,国际上对 MIMIC 数据集的研究正处于上升时期。

4.2.2 发文国家及机构分布

对研究领域内发文国家和机构进行分析,可以了解各国及研究机构在该领域内的合作关系、科研力量、研究水平等情况。

(1)发文国家分布

MIMIC 数据集的开源性受到了各国研究者的欢迎,共有美国、中国、英国等 20 个国家的研究人员对 MIMIC 数据集进行研究,表 5 显示了国际上研究力量排名前 10 的国家。由表 5 可见,美国的发文量遥遥领先其他国家,共 63 篇,占文献总量的 43.45%,其次是中国、葡萄牙、法国以及英国,分别是 18 篇、9 篇、8 篇、

8篇。最后印度、伊朗等国家也有一定的发文量。总体来说，美国的发文数量最多，科研力量最强，这和MIMIC数据集是由美国贝斯以色列迪康医疗中心开放有一定的关系。同时在统计过程中，发现美国和其他国家也有着广泛的合作。

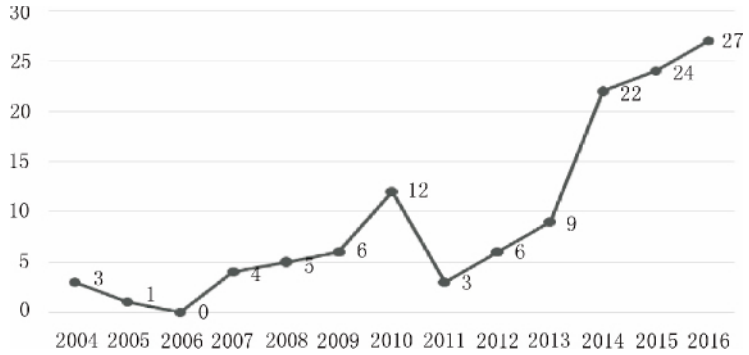

图1　关于MIMIC数据集研究年发文量

表5　关于MIMIC数据集研究各国发文量统计表

国家	发文量	占比(%)
美国	63	43.45
中国	18	12.41
葡萄牙	9	6.21
法国	8	5.52
英国	8	5.52
印度	6	4.14
伊朗	4	2.76
加拿大	4	2.76
德国	3	2.07

(2)发文机构分布

利用MIMIC数据集进行科学研究的机构包括四类，分别是大学、医疗机构、医学研究所和医疗公司，统计每类中主要代表的发文量。各机构利用MIMIC数据集进行科学研究的发文量统计表如表6所示，其中高校的发文量最多，共66篇，占比52.80%，主要代表高校有美国哈佛大学、麻省理工大学、中国的浙江大学、葡萄牙的里斯本大学和英国的牛津大学，其中哈佛大学和麻省理工大学发文量较多，分别是17.6%、21.6%；其次是医学研究所，共发文7篇，占比5.60%，主要代表机构是中国科学院和法国国家卫生医学研究所；同时有些医疗公司也利用MIMIC数据集进行科学研究，主要代表是美国的Dascena公司和飞利浦医疗保健中心；最后是医疗机构，发文量最少，占比3.20%，主要代表机构是贝斯以色列迪康医疗中心。这说明MIMIC数据集在学术界和工业界都引起了研究者的关注，此外，关于MIMIC数据集研究的主要研究力量是高校，其次是医学研究所和医疗公司。

表6　关于MIMIC数据集研究各研究机构发文量统计表

研究机构	主要代表	发文量	占比(%)	合计(%)
大学	麻省理工大学	27	21.60	52.80
	哈佛大学	22	17.60	
	里斯本大学	6	4.80	
	牛津大学	4	3.20	
	浙江大学	7	5.60	
医疗机构	贝斯以色列迪康医疗中心	3	2.40	3.20
	济南第四医院	1	0.80	
医学研究所	法国国家卫生医学研究所	3	2.40	5.60
	中国科学院	4	3.20	
医疗公司	Dascena公司	4	3.20	5.60
	飞利浦医疗保健中心	3	2.40	

4.2.3　研究主题分布

研究某个领域的主题分布可以一定程度上了解该领域内的研究热点，表7是关于MIMIC数据集研究主题分布的统计情况。由表7可知，研究主题主要集中在7个方面，分别是：①ICU病人预后及死亡率预测；②影响ICU病人预后或者死亡率的因素；③ICU病人基本生命体征信息的研究；④探究某些因素是否是影响某些疾病的影响因子；⑤关于MIMIC数据集介绍和数据处理方法的研究；⑥预测某种疾病的发病率或死亡率；⑦其他。

ICU 病人预后及死亡率预测主题，主要是建立模型对 ICU 病人预后结果和死亡率进行预测，如 Lehman 等基于 SVAR 框架，提出了一种基于生命体征时间动力学的方法，用于预测和跟踪患者在住院期间存活的倾向，以及他们的 28 天存活率[18]；Fuchs 等人利用回顾性队列研究方法，从病人特点、疾病严重程度、护理强度和死亡率几个方面分析老年重症监护病房住院率和死亡率的变化趋势，研究结果表明老年人 ICU 入院疾病严重程度的降低导致 ICU 入住率的升高，但和 ICU 死亡率无关[7]。

影响 ICU 病人预后或者死亡率的因素的主题研究，主要是建立模型探究哪些因素影响 ICU 病人预后结果或者死亡率，如 Mayaud 等人利用遗传算法从脓毒症患者在低血压期间的动态变量中找出影响医院死亡率预测的潜在变量，开发一种新的死亡率预测方法，比现有预测方法有更高的辨识度，为患者更好地提供预后[5]。

ICU 病人基本生命体征信息研究主要是利用算法研究病人的基本生命体征信息，如血压、心率以及呼吸速率等，如 Shamim 等人利用融合算法提出了一个修改的卡尔曼滤波器(KF)框架的应用，用于数据融合，以估计来自多个生理源的呼吸率，能够更好地估计病人的呼吸速率[19]。

探究某些因素是否是影响某些疾病的影响因子的主题研究，主要是利用模型分析某种因子是否是某些疾病的影响因子，如 Saeed 利用多变量逻辑回归以发现低血压和急性肾损伤(AKI)之间的相关性，发现在危重疾病情况下，低血压可能与急性肾损伤(AKI)有关[14]。

关于 MIMIC 数据集介绍和数据处理方法的研究主要是介绍 MIMIC 数据集的基本情况和一些数据处理方法，如 Clifford 等人以鲁棒参数提取为例，提供了处理 ICU 复杂的、不规则数据的方法，包括数据的收集、测量、转录、提取以及降噪等[20]。

预测某种疾病的发病率或死亡率的主题研究，主要是建立模型预测某种疾病的发病率或死亡率，如 Celi 等人提出了一种新的患有急性肾损伤的 ICU 患者死亡率预测的模型，该模型比 SAPS 系统预测更准确，表明定制建模可能提供更准确的预测[10]。

其他类主题研究范围较广，包括探究不同测量方法的效果差异等，如 Clifford 等比较临床医生记录的有创血压和监测设备自动记录有创血压与患者未来低血压的相关度，结果表明自动化记录血压的方法比人工护理人员记录的血压方法具有更高的灵敏度和特异性[21]。

由表 7 可知，ICU 病人预后或者死亡率预测和 ICU 病人生命体征信息的研究这 2 个主题的文献量很多，分别是 20 篇、19 篇。其次是关于 MIMIC 数据集介绍或者数据处理方法的研究、预测某种疾病的发病率这 2 个主题的文献量较多，都是 16 篇。探究某些因素是否是影响某些疾病的影响因子和影响 ICU 病人预后或死亡率的因素这两个主题的文献量较少，分别是 11 篇、9 篇。其他类主题的文献量最多，共 21 篇。从总体看来，关于 MIMIC 数据集研究的热点主要集中在对 ICU 病人预后和死亡率的预测和对 ICU 病人基本生命体征信息的研究上。由于 ICU 病房资源非常宝贵，因此提高 ICU 病人预后和死亡率预测的正确率，对合理分配 ICU 的医疗资源有重要意义。

表 7　关于 MIMIC 数据集研究主题分布统计表

主题	发文量	占比(%)
ICU 病人预后及死亡率预测	20	17.86
影响 ICU 病人预后或死亡率的因素	9	8.04
ICU 病人基本生命体征信息的研究	19	16.96
探究某些因素是否是影响某些疾病的影响因子	11	9.82
关于 MIMIC 数据集介绍或数据处理方法的研究	16	14.29
预测某种疾病的发病率或死亡率	16	14.29
其他	21	18.75

4.2.4　研究方法分布

为了较为全面地了解关于 MIMIC 数据集研究的情况，笔者统计了采用 MIMIC 数据集进行研究的文献使用的研究方法，由于有些研究没有具体的数据处理方法，因此选择有明确数

据处理方法的30篇文献进行统计分析，如表8所示。研究方法包括数据采集方法和数据处理方法两大类，数据采集方法包括根据前瞻性队列研究采集数据和根据回顾性队列研究采集数据，前瞻性临床研究是研究者根据选题和设计的要求而进行的研究，按设计要求详细记录临床资料，通过对这些资料的整理、归纳、统计、分析，得出某一结论。回顾性队列研究是从以往临床工作积累的病例资料中，选择某一时期同类临床资料进行整理、分析，以从中总结经验、找出规律、指导实践的研究[21]。由于回顾性队列研究是选取已经存在的数据进行分析研究，因此是主要的采集数据的方法，占比86.67%，而只有极少一部分研究采用前瞻性队列研究方法。

表8　关于MIMIC数据集主要研究方法统计表

方法类别	主要研究方法	频次	占比(%)
数据采集方法	回顾性队列研究	26	86.67
	前瞻性队列研究	4	13.33
数据处理方法	回归分析	7	23.33
	相关分析	5	16.67
	可视化	2	6.67
	机器学习算法	10	33.33
	其他	6	20.00

关于MIMIC数据集研究文献中用的数据处理方法主要包括回归分析、相关分析、可视化方法、机器学习算法五大类，其中机器学习算法应用最广泛，占比33.33%，这主要和机器学习算法常用于处理大规模数据有关，机器学习算法常用于研究ICU病人预后及死亡率预测主题。相关分析和回归分析应用也是主要的研究方法，主要用于影响ICU病人预后或者死亡率的因素和探究某些因素是否是影响某些疾病的影响因子这两大主题的研究。在其他方法中主要是研究者自建模型来进行数据处理。应用最少的可视化方法主要用于MIMIC数据集可视化展示。

5　结论

MIMIC医疗数据集自从免费开放以来，国际上关于该数据集的研究文献持续增多，因此系统梳理MIMIC数据集的内容以及相关研究具有重要意义。本文通过整理MIMIC数据集内容和对关于MIMIC数据集研究文献的基本情况进行统计与内容剖析，主要得出以下结论与建议：

(1)MIMIC数据集包含四类信息，分别是病人基本信息及转移信息、病人医院门诊治疗的相关信息、病人在ICU治疗的相关信息和辅助信息。其中大部分信息在相关研究中被采用，然而部分信息尚未被研究利用，且其往往与专门领域的深度挖掘有关，如微生物检测信息表和排泄事件表可用于生物信息领域的深度挖掘，医疗过程事件表和病人注射事件表涉及药物不良反应深度挖掘，这说明该数据集的研究深度与广度不足，有待结合相关领域知识与外部数据，进一步挖掘分析。

(2)涉足MIMIC数据集挖掘研究的国家众多，高校是主要的依托机构，且相关研究正处于上升期，其中美国占据了主导地位，我国也具有较强的影响力。但是，我国高校与医疗公司在MIMIC数据集研究方面远远落后于美国，这应该引起管理部门与相关行业科研人员的重视。

(3)MIMIC数据集的研究主题较集中，研究热点是ICU病人预后与死亡率预测及其影响因素。但是，MIMIC数据集是拥有丰富类型医疗数据的真实大数据集，研究者可以有更多的研究选题，如从MIMIC数据集的大量临床文本数据NOTEEVENTS表中，挖掘并构建知识库，可以有效辅助临床决策；针对某种具体疾病进行影响因素分析和提前预测分析，从而更好地提醒人们如何预防该疾病的发病；根据病人的病情和基本生理特征情况，预测病情发展进程，为患者提供更好的预后并合理分配医疗资源等。

(4)MIMIC数据集研究采用的数据采集方法主要是回顾性研究方法，数据处理方法上一般是应用通用挖掘方法，主要有基于朴素贝叶斯或K-Means的机器学习、回归分析与相关分析。因此，MIMIC数据集的分析挖掘方法，在针对性与挖掘效能上有较大的发展空间，结合医学语义知识及一些新颖高效的挖掘方法，如医学关联数据、大数据深度学习及迁移学习，进行MIMIC数据集的挖掘模型研究，也是应该关注的方向。

参考文献

[1] **马锡坤，杨国斌，于京杰. 国内电子病历发展与应用现状分析**[J]. **计算机应用与软件**，2015，32(1)：10-12

[2] James M，Michael C，Brad B，et al. Big data：The next frontier for innovation，competition，and productivity[EB/OL].[2017-05-17]. http://webanalisten.nl/wp-content/uploads/2011/05/MGI_big_data_full_report.pdf

[3] **王剑，张政波，王卫东，等. 基于重症监护数据库** MIMIC-Ⅱ **的临床数据挖掘研究**[J]. **中国医疗器械杂志**，2014，38(6)：402-406

[4] Dejam A，Malley B E，Feng M，et al. The effect of age and clinical circumstances on the outcome of red blood cell transfusion in critically ill patients[J]. Critical Care，2014，18(4)：1-9

[5] Mayaud L，Lai P S，Clifford G D，et al. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension[J]. Critical Care Medicine，2013，41(4)：954-62

[6] Lee J，Louw E，Niemi M，et al. Association between fluid balance and survival in critically ill patients[J]. Journal of Internal Medicine，2015，277(4)：468-477

[7] Fuchs L，Novack V，Mclennan S，et al. Trends in severity of illness on ICU admission and mortality among the elderly[J]. Plos One，2014，9(4)：e93234

[8] Moskowitz A，Lee J，Donnino M W，et al. The association between admission magnesium concentrations and lactic acidosis in critical illness[J]. Journal of Intensive Care Medicine，2014，301(6)：444 - 451

[9] Goldstein I，Özlem Uzuner. Specializing for predicting obesity and its co-morbidities[J]. Journal of Biomedical Informatics，2009，42(5)：873-886

[10] Celi L A G，Tang R J，Villarroel M C，et al. A clinical database-driven approach to decision support：predicting mortality among patients with acute kidney injury[J]. Journal of Healthcare Engineering，2011，2(1)：97-109

[11] Lee J，Kothari R，Ladapo J A，et al. Interrogating a clinical database to study treatment of hypotension in the critically ill[J]. BMJ Open，2012，2(3)：1-10

[12] Ghassemi M M，Richter S E，Eche I M，et al. A data-driven approach to optimized medication dosing：A focus on heparin[J]. Intensive Care Medicine，2014，40(9)：1332-1339

[13] Hunziker S，Celi L A，Lee J，et al. Red cell distribution width improves the simplified acute physiology score for risk prediction in unselected critically ill patients[J]. Critical Care，2012，16(3)：1-8

[14] Lehman L，Saeed M，Moody G，et al. Hypotension as a risk factor for acute kidney injury in ICU patients[C]// Computing in Cardiology，Belfast，UK，2010：1095-1098

[15] Lehman L H，Saeed M，Talmor D，et al. Methods of blood pressure measurement in the ICU[J]. Critical Care Medicine，2013，41(1)：34-40

[16] Hug C W，Clifford G D，Reisner A T. Clinician blood pressure documentation of stable intensive care patients：An intelligent archiving agent has a higher association with future hypotension[J]. Critical Care Medicine，2011，39(5)：1006-1033

[17] Neamatullah I，Douglass M M，Lehman L W，et al. Automated de-identification of free-text medical records[J]. BMC Medical Informatics and Decision Making，2008，8(1)：1-17

[18] Lehman L W，Adams R P，Mayaud L，et al. A physiological time series dynamics-based approach to patient monitoring and outcome prediction[J]. IEEE Journal of Biomedical & Health Informatics，2015，19(3)：1068-1087

[19] Nemati S，Malhotra A，Clifford G D. Data fusion for improved respiration rate estimation.[J]. Eurasip Journal on Advances in Signal Processing，2010(1)：e926305

[20] Clifford G D，Long W J，Moody G B，et al. Robust parameter extraction for decision support using multimodal intensive care data[J]. Philosophical Transactions，2008，367(1887)：411-429

[21] Hug C W，Clifford G D，Reisner A T. Clinician blood pressure documentation of stable intensive care patients：an intelligent archiving agent has a higher association with future hypotension[J]. Critical Care Medicine，2011，39(5)：1006-1026

(**收稿日期**：2017-06-23)