国际医学信息学领域大数据研究热点分析

黄 鹂 曹东维

（南京大学医学院附属鼓楼医院 南京 210008）

〔摘要〕 基于 Web of Science 从发文年代、国家/地区和机构、核心作者及主要期刊分布几方面分析医学信息学学科大数据研究的现状和进展，借助软件对关键词进行聚类分析，发现研究热点主题主要集中在临床决策支持系统、临床研究数据管理、电子健康档案、转化生物信息学和遗传流行病学等方面。

〔关键词〕 大数据；医学信息学；聚类分析；研究热点

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.04.001

Analysis of Big Data Study Hotspots in the International Medical Informatics Field *HUANG Li, CAO Dong－wei, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing 210008, China*

〔**Abstract**〕 Based on Web of Science, the paper analyzes the current situation and progress of big data study of the medical informatics subject from the aspects like publication time, countries/regions and institutions, core authors and distribution of major periodicals. It carries out cluster analysis of keywords with the help of software and finds out that subjects of study hotspots mainly focus on such aspects as Clinical Decision Support System (CDSS), clinical research data management, Electronic Healthcare Records (EHR), translational bioinformatics and genetic epidemiology, etc.

〔**Keywords**〕 Big data; Medical informatics; Cluster analysis; Study hotspots

1 引言

在生物医学信息学领域大数据是一种新的范式和生态系统，它将基于病历的研究转化为大规模的数据驱动研究。首先，电子健康记录、医学影像设备的使用使海量患者数据以电子格式被收集和存储；其次，基因组学、蛋白质组学、代谢组学等测序系统与电子健康档案系统、临床实验结果、医疗传感器等共同产生各种数据类型和结构；最后，大数据技术的发展如人工智能、Hadoop 和数据挖掘工具为生物医学研究人员提供新的模式。在此背景下生物医学科学家面临着存储、管理和分析海量数据集的挑战，大数据的特点需要强大而新颖的技术来提取有用的信息并实现广泛的医疗保健解决方案[1]。

2 数据来源

本文的数据源为 Web of Science 核心合集，检索式为主题：("big data") OR 主题：("data－driven")，时间跨度为 1986－2017 年。类别选择 Medical Informatics，精炼得到医学信息学学科研究领域以"big data" /"data－driven" 为主题的 1 448 条记录，对其进行分析，在此基础上结合文献和词频对热

〔修回日期〕 2018－04－20

〔作者简介〕 黄鹂，馆员，发表论文 5 篇。

点关键词进行筛选，尝试揭示国际医学信息学大数据研究热点。

3 结果与分析

3.1 发文年代

年度发文量是文献量在时间节点上的映射，也是研究热度随时间推移的表现。将 1 448 条记录导入 Hiscite 软件进行统计，见图 1。可以发现医学信息学领域的大数据研究总体发展可以分为以下 3 个阶段：1991－1999 年为萌芽期，2000－2011 年为稳定递增期、2012－2017 年为高速发展期。在跨度为 9 年的萌芽时期，年度文献量在 20 篇左右并缓慢递增，最早的一篇论文为 Sittig D F 于 1988 年发表在《医疗决策》上的“使用数据驱动的计算机化患者监护仪来降低成本并提高患者护理质量”一文，当时已利用数据驱动的计算机管理系统来降低成本和提高护理质量，认识到大数据对临床决策的作用。在跨度为 7 年的稳定递增期，年度文献量逐年稳定递增，发文量虽有波动但幅度不大。2012 年以来（数据更新至 2017 年 12 月 31 日）是研究热点时期，年文献量增长迅速，在 2016 年文献量达到顶峰，这说明随着大数据科学的发展，医学信息学领域的相关研究越来越多。2017 年发文量有所下降，但这并不意味着研究的停滞，一方面可能由于数据收录尚不完善，另一方面医学信息学领域的大数据研究正面临着新的突破，如数据隐私和数据安全、核心理论等其他学术问题制约研究的深化和拓展，使研究出现阶段性下降趋势，符合学科发展的规律。

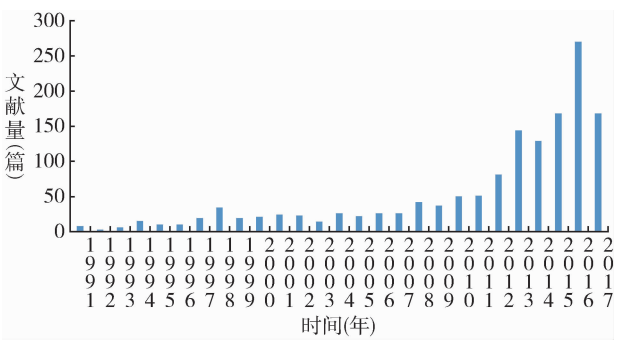

图 1 医学信息学领域大数据发文年代分布

3.2 发文国家/地区和机构

国家/地区的分布情况可反映一个研究领域的主要研究力量空间分布态势。通过对 1 448 篇文献进行统计分析，可知研究力量广泛分布在 76 个国家/地区所属的 1 842 个研究机构中。从统计结果中可以发现以下特征：（1）发文量大于 10 篇的国家/地区有 29 个，研究机构 31 个。（2）从发文数量看，美国发文最多（594 篇），是排名第 2 英国的 4.86 倍，其次为德国、中国、加拿大、西班牙、澳大利亚等国家；发文量大于 26 篇的国家/地区有 19 个，集中分布在欧洲和亚洲地区，见图 2。（3）发文量前 10 的研究机构全部属于美国，排名第 1 的是发文 30 篇的美国密歇根大学。研究机构除梅奥诊所，其余是大学中的医学信息学院和计算机学院，可以看出医学信息学领域大数据研究内容跨学科特征明显且实用性较强，见图 3。

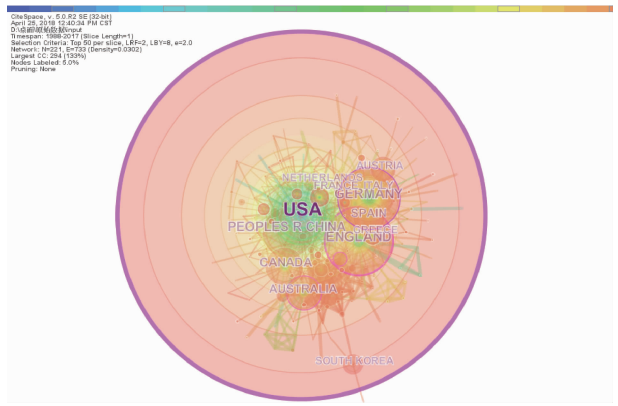

图 2 医学信息学领域大数据发文国家/地区

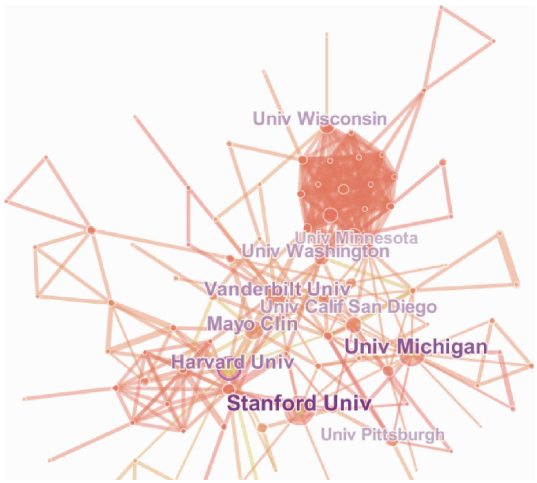

图 3 医学信息学领域大数据发文机构

3.3 核心作者及主要期刊

利用 Hiscite 软件对作者进行分析，最高产作者是美国圣地亚哥大学医学院的 Ohno – Machado Lucila，共发表 11 篇文献，6 篇是第 1 作者，被引频次达到 76，作者对医学大数据挖掘、共享、算法和工具等进行研究。发文量前 10 的作者总被引频次均在 15 以上，可被认为是本领域的核心作者，其中两人来自梅奥诊所，分别是 Chute CG 和 Pathak J，8 位来自大学信息和医学学院的作者。发文量前 10 的期刊均是 SCI 收录期刊，发文量 40 篇以上。发文最多的是《美国医学信息学会杂志》（*Journal of the American Medical Informatics Association*）（146 篇），该杂志提供最新信息帮助医师、信息学家、科学家、护士及其他保健护理专业人员在患者护理、教学、研究及保健管理中开发与利用信息学技术。发文量第 2 的是《生物医学信息学杂志》（*Journal of Biomedical Informatics*）（103 篇），发表有关生物医学情报学方法论与计算机应用研究的文章。其他杂志依次为《医学互联网研究杂志》（73 篇）、《生物医学计算机方法与程序》（56 篇）、《国际医学信息科学杂志》（56 篇）、《医学统计学》（50 篇）等，发文期刊主要集中在医学信息学、计算机科学领域。

4 研究热点聚类分析

4.1 概述

利用 CiteSpace 软件对选取的 1 448 篇文献进行关键词提取和词频统计，围绕大数据、数据驱动的医学信息学领域产生一系列关键词如“电子健康档案”、“系统”、“大数据”等，但单纯以词频衡量一个词的重要性不够全面，通过 TF – IDF 加权生成高相关性的关键词，如“临床决策支持系统”、“数据驱动计划”、“社交媒体”、“药物基因组学研究”等。通过对 1 448 篇文献的热点关键词进行聚类分析，得到医学信息学领域大数据研究热点聚类，见图 4。结合高被引论文、核心作者论文等相关文献，得到医学信息学领域大数据的 5 个热点研究主题。

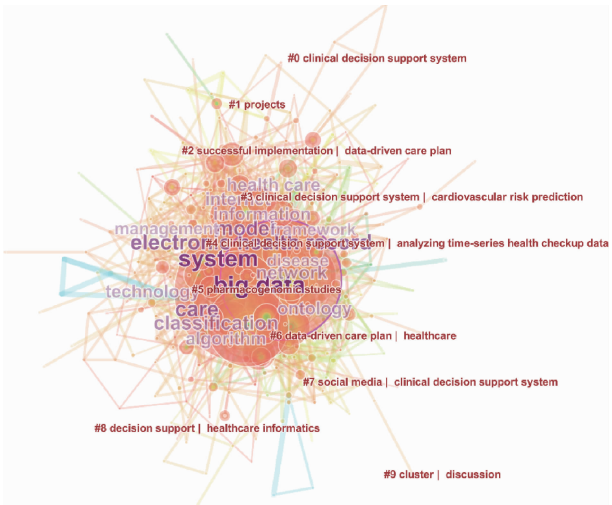

图 4 医学信息学领域大数据研究热点聚类

4.2 临床决策支持系统

美国医学信息学会对临床决策支持系统定义是为医务工作者、患者或任何个人提供知识、特定个体或人群信息，在恰当的时间智能化地过滤和表达信息，提供更好的健康、诊疗和公共卫生服务；或者是在正确的时间对正确的对象提供正确的信息，这有别于人工智能和专家系统[2]。大数据及生物信息学领域的跨国公司正致力于分析大型生物信息数据集来发现医学的新进展，对疾病的诊断、理解和治疗将产生深远影响[3]。临床决策支持尝试通过警报、模板和预测评分系统加强最佳实践，应用决策系统后临床医生可以通过关键字搜索并输入命令以触发随后的临床行为，如搜索“肺炎”就能找到相关命令的模板（如血培养、抗生素、胸部 X 光片等）。Jose′ M 设计人造神经网络工具，用于分析数据和待预测信息之间存在的复杂非线性相互作用的数据集，能对癌症术后患者的临床预测提供决策支持。作者提出一种预测乳腺癌复发预后的决策支持工具，结合 TDIDT 算法，选择与乳腺癌预后最相关的因子，结合西班牙马拉加大学医学院的临床病理数据进行分析，建立一个乳腺癌术后复发预后模型，供临床医生用来搜索寻找预后因素中的细微模式的大数据集，进一步帮助患者选择合适的辅助治疗[4]。

4.3 临床研究数据管理

医院常规收集的数据是复杂的异质数据，分散在多个医院信息系统中，使用常规数据来构建数据驱动的临床路径，描述过程和趋势，能更好地了解疾病，发掘隐藏的模式，揭示医院中可能无法获得或忽略的患者和疾病的信息，改善服务和成本。这种数据的范围和用途依赖于其质量，传统的算法不能很好地处理非结构化的过程或数据，也不能产生具有临床意义的可视化。临床路径也被称为护理或关键路径，已被引入到卫生保健系统以提高护理效率，同时维持或改善其质量，临床路径由多学科团队开发并依赖于医院信息系统和电子健康记录[5]。Joao H 探讨常规医院数据如何应用于发展数据驱动的路径以描述患者的护理过程以及在生物医学研究中的潜在用途。作者从一所大型英国医院提取与前列腺癌患者有关的数据，辅以当地癌症登记处的信息，在 1904 年的患者和专家知识库中建立数据驱动的路径，其中包含前列腺癌生物标志物的规则，用于评估特定临床研究路径的完整性和效用。构建质量评估和可视化的患者路径，对复杂的以患者为中心的临床信息进行概括、可视化和查询，对质量指标和维度进行计算，这些方法可在其他环境中构建疾病途径的数据收集、演示和质量评估，是医院利用医学大数据的重要进展[6]。

4.4 电子健康档案

它是人们在健康相关活动中直接形成的具有保存备查价值的电子化历史记录，存储于计算机中，是面向个人提供服务且具有安全保密性能的终身个人健康档案。电子健康档案的研究起始于 20 世纪 90 年代中后期[7]。其采用使研究人员更容易获取和汇总临床数据，电子健康档案数据的 2 次使用是一个有潜力的研究领域，人们越来越关注获取临床护理过程中的数据并进行研究，对数据的质量进行评估，建立一致的数据质量维度，建立系统的数据研究方法，开发和分享评估数据质量的最佳实践[8]。临床决策支持需要历史电子病历数据支持，通过对历史数据中临床项目要素（药物、实验室、护理指令、入院诊断、ICD9 代码等）的评估，从过去的临床行为中通过算法挖掘临床项目和时间的关联规则，可对未来的临床行为进行预测，最好是相近日期的数据。这些数据对研究慢性疾病、不良反应、疾病预防和流行病学来说有着重要意义[9]。

4.5 转化生物信息学和遗传流行病学

生物医学信息学在大科学和大数据中占有重要的地位，对大数据的信息学分析已经被整理到床边并影响患者的结果[10]。转化生物信息学是医学信息学研究的新领域，它将生物信息数据和临床知识结合在一起，分析和挖掘数据，探索基因和疾病之间的关系。遗传流行病学搜集数据并进行注释，使用工具对人类 DNA 进行可视化和检索[11]。医学/基因组学研究是大数据技术应用的理想案例，应用领先的大数据解决技术（如 Hadoop、CloudBurst、Crossbow、CloudAligner、Myrna 等研究工具）对基因组进行数据分析、比对并进行排序。领先的数据分析软件 Apache、Hadoop 和服务提供商 Cloudera 与美国西奈山医学院的基因组学生物研究所合作，应用大数据技术来诊断和治疗疾病，研究领域包括人类和细菌基因组的分析，研究机体正常和疾病状态的代谢途径，用于治疗疾病的分子结构和功能等，致力于尖端领域的研究，形成强大的合作[3]。

4.6 大数据分析和挖掘

当前数据量爆炸性增长，尤其是医疗行业产生大量的数据，包括临床记录、医学图像、基因组数据、健康行为、临床决策支持、疾病监测和公共健康管理等，充分利用这些医疗大数据不仅要建立数据挖掘系统，还要为数据挖掘和医学信息社区之间的跨学科研究建立桥梁。医学信息学是一个自然的框架，可以在决策环境中正确有效地应用数据分析和数据挖掘方法。未来有必要保持该领域的包容性，促进研究人员之间的数据和方法共享，更好地开发和利用医学大数据[12]。Jake Luo 将大数据在生物医学研究和卫生保健的应用分为 4 类：生物信息学、临床信息学、影像信息学和公共卫生信息学。具体而言在生物信息学中分析分子水平上生物系统

的变化，高通量实验有助于研究新的全基因组关联疾病，并且利用临床信息学，临床领域将从大量收集的患者数据中获益和决策，影像信息学快速与云平台集成，以共享医学图像数据和工作流程，公共健康信息利用大数据预测和监测如埃博拉病毒等传染病爆发[13]。Aisling 概述云计算和大数据技术，讨论如何利用这些专业知识来处理生物学的大数据集，特别是大数据技术如 Apache Hadoop 项目将提供分布式并行化的数据处理和 PB 级规模的数据集分析，概述其在生物信息学界的使用情况[3]。

4.7 社交媒体大数据

随着 Web2.0 普及，在线社交媒体应用如 Facebook、Twitter、微博等不断涌现，产生海量的社交媒体大数据。在社交媒体中用户既是数据接收者，也是数据生产者。社交媒体大数据是巨大的资源，Twitter、Instagram 等知名社交媒体网站产生大量与药物有关的用户生成的内容和用户交互（如安眠药、抗抑郁药等的滥用）主题的社交网络大数据，可以了解、监测和干预药物滥用和成瘾问题。一些人不仅在社交媒体上分享和交流关于药物使用的经验，还通过网络寻求具有类似成瘾问题群体的社会支持。SJ Kim 建立多维框架对社交媒体大数据进行分析，讨论药物使用相关交流的流行和社会影响、用户特征、通信特性、机制和预测因素、道德领域等，以促进对处方药使用问题的全国性危机的研究[14]。Young 介绍近期在生物信息学、数字流行病学和疾病建模领域中的工作，描述如何将其应用于艾滋病毒预防，提出在实施移动技术大数据方法预防艾滋病毒之前需要解决的问题[15]。

4.8 数据隐私和安全

医学数据不可避免地涉及患者的隐私信息，数据隐私和安全是医学信息学研究的主要内容，为研究人群健康状况、疾病病因和医学治疗的有效性等，临床研究需要访问大量患者数据库，由于健康研究涉及受保护的人体数据，只能在适当的监管和隐私保护的框架内进行。Wolfgang 开发标准的患者数据隐私框架模型，该模型允许分析数据隐私和相关性问题，以结构化方式对患者数据进行研究，提供一个框架来兼容数据流，找出薄弱环节并改进。大数据时代保护医疗和基因组数据以及不断变化的隐私环境面临非常大的挑战，尽管云计算作为处理大数据集的方法得到支持，但是数据安全问题阻碍其在生命科学商业领域的广泛应用。亚马逊、谷歌等跨国公司已经意识到其中巨大的商机，开发一系列的系统和数据工具，利用大数据技术来产生经济效益的动力越来越强，因此数据隐私和安全尤为重要。2016 年 4 月 14 日欧洲议会投票通过商讨 4 年的“一般数据保护条例”，新条例的通过意味着欧盟对个人信息保护及其监管达到前所未有的高度，堪称史上最严格的数据保护条例；美国颁布个人信息隐私保护法案；德国在保护个人信息方面的立法已有几十年历史，现行的《联邦数据保护法》于 2009 年修改并生效，约束范围包括互联网等电子通信领域，旨在防止因个人信息泄露导致的侵犯隐私行为。中国的数据保护法或个人信息保护法尚未完善推出，在数据隐私和安全保护方面需要加强立法。

5 结语

当前大数据技术正迅速应用于生物医学和医疗保健领域，在新的行业分析报告中麦肯锡公司预测医疗领域的大数据分析将在美国每年节省超过3 000亿美元的保健费用，大数据应用在生物医学领域未来的发展具有可预见的作用，其依赖于新数据标准的推进、相关研究和技术的发展、研究机构和公司的合作以及强大的政府激励机制。医学信息学领域大数据的发展将随着研究的深入不断推进，对计算机科学、生物医学、医学信息学等多学科产生深远的影响。在国内大数据的应用还有待开发，一些数据公司如医渡云等开始与医院开展合作，建立区域医疗数据中心，应用人工智能技术对医疗数据进行集成、挖掘和利用，这些数据为展开医学信息学领域的大数据研究提供了生物信息数据集。在医学信息学研究中应利用自身优势，对现有电子健康档案和临床研究数据进行多渠道挖掘，分析生物信息数

据集为临床决策提供参考和支持，通过大数据分析技术将生物信息数据和临床知识结合在一起进行实践探索。同时大数据研究从基础研究转向应用研究需要建立跨学科研究团队进行多学科合作，医学信息学研究人员将充当技术架构师的角色，为大数据研究整合不同的方法和工具。总之医学信息学领域的大数据研究前景是美好的，挑战也一并存在。由于数据类型和标准的不同，各种数据库的整合和利用存在困难，同时数据安全和隐私问题应该引起重视，大数据在医学信息学领域的研究尚处于起步阶段，尽管功能强大、意义深远，但是还有很长的路要走。

参考文献

1 Jake Luo, Min Wu, Deepika Gopukumar, et al . Big Data Application in Biomedical Research and Health Care: a literature review [J]. Biomedical Informatics Insights, 2016, (8): 1 - 10.

2 董建成. 医学信息学概论 [M]. 北京：人民卫生出版社，2010：260 - 263.

3 A O'Driscoll, J Daugelaite , RD Sleator. 'Big data', Hadoop and Cloud Computing in Genomics [J]. Journal of biomedical informatics, 2013, 46 (5): 774 - 781.

4 Jose' M Jerez - Aragone's, Jose' A Go'mez - Ruiza, Gonzalo Ramos - Jime'neza, et al. A Combined Neural Network and Decision Trees Model for Prognosis of Breast Cancer Relapse [J]. Artificial Intelligence in Medicine, 2003, 27 (1) : 45 - 63.

5 Vanhaecht K , Panella M, RV Zelm, et al. An Overview on the History and Concept of Care Pathways as Complex Interventions [J]. International Journal of Care Pathways, 2010, 14 (3): 117 - 123.

6 Joao H Bettencourt - Silva, J Clark, CS Cooper, et al. Building Data - Driven Pathways From Routinely Collected Hospital Data: a case study on prostate cancer [J]. JMIR medical informatics, 2015, 3 (3): e26.

7 罗忠宁. 医学信息学 [M]. 兰州：兰州大学出版社，2012：149.

8 Nicole Gray Weiskopf, C Weng. Methods and Dimensions of Electronic Health Record Data Quality Assessment: enabling reuse for clinical research [J]. Journal of the American Medical Informatics Association, 2013, 20 (1): 144 - 151.

9 JH Chen, M Alagappan, MK Goldstein, et al. Decaying Relevance of Clinical Data Towards Future Decisions Indata - driven Inpatient Clinical Order Sets [J]. International Journal of Medical Informatics, 2017, (102): 71 - 79.

10 Lucila Ohno - Machado. Big Science, Big Data, and a Big Role for Biomedical Informatics [J]. Journal of the American Medical Informatics Association, 2012, 19 (1): e1.

11 Paul A Harris, RobertTaylor, RobertThielke, et al. Research Electronic Data Capture (REDCap) - a metadata - driven methodology and workflow process for providing translational research informatics support [J]. Journal of Biomedical Informatics, 2009, 42 (2) : 377 - 381.

12 R Bellazzi, M Diomidous, IN Sarkar et al. Data Analysis and Data Mining: current issues in biomedical informatics [J]. Methods of information in Medicine, 2011, 50 (6): 536 - 544.

13 Jake Luo, Min Wu, Deepika Gopukumar, et al. Big Data Application in Biomedical Research and Health Care: a literature review [J]. Biomedical Informatics Insights, 2016, (8): 1 - 10.

14 SJ Kim, LA Marsch , JT Hancock, et al. Scaling Up Research on Drug Abuse and Addiction Through Social Media Big Data [J]. Journal of Medical Internet Research, 2017, 19 (10): e353.

15 Y Zhang, M Huo, J Zhou, et al. PKSolver: an add - in program for pharmacokinetic and pharmacodynamic data analysis in Microsoft Excel [J]. Computer Methods and Programs in Biomedicine, 2010, 99 (3): 306 - 314.