

## 大数据挖掘促进精准医学发展\*

任思冲, 周海琴, 彭 萍<sup>△</sup>

(成都市龙泉驿区第一人民医院, 四川成都 610100)

DOI: 10.3969/j.issn.1673-4130.2015.23.066

文献标识码: B

文章编号: 1673-4130(2015)23-3499-03

随着互联网、社交网络、物联网、云计算等新一代信息技术的应用和推广,大数据时代应运而生。大数据库中含有丰富的信息资源,其潜在价值越来越大,数据信息的提取和挖掘能力成为了大数据时代的发展战略需求,需要利用多种数据挖掘工具来开发其潜在的有效价值,以促进多学科的发展和更新。医学的发展经历了传统医学和循证医学的阶段,如今发展到以分子生物学为本质出发点,对疾病精准的预防、诊断和治疗的精准医学时代。作为现代医疗模式的革命和创新——精准医学及精准医疗时代的到来,挖掘和整合涵盖流行病学、预防医学、临床医学、康复医学、卫生经济学和医学分子生物学等多学科领域的大数据将成为精准医学发展的首要任务之一,通过大数据挖掘分析技术提取有效的价值,可以指导和制订出适合每位患者的精准的个体化预防和治疗方案,以期达到治疗效益最大化和医疗资源配置最优化。

## 1 大数据定义

大数据是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合,大部分专业人士认为它是继云计算、物联网之后信息技术产业又一次颠覆性的技术变革产物。它有别于以往的“海量数据”,大数据不仅在于数据量大,更在于其具有数据类型繁多、价值密度低、处理速度快等属性特征<sup>[1]</sup>。(1)数据量大:我国医疗卫生服务和各种医疗卫生信息系统产生了巨量数据,以每个 CT 图像约 150 MB 的数据、每个基因组序列文件约为 750 MB、每个标准病理图接近 5 GB 计算,乘以我国人口数量和平均寿命,那么每个社区医院或中等规模制药企业均可以生成和累积达数个 TB 甚至数个 PB 级的结构化和非结构化数据。(2)类型繁多:医疗数据类型复杂,不仅限于电子病历中患者的基本数据、输入转出数据等结构化数据,还包括医学影像数据,临床实验室检测数据及互联网中存在的医学数据等海量的半结构化和非结构化数据。(3)价值密度低:目前大数据的价值存在稀疏的特点,价值密度的高低与数据总量的大小成反比,因此,需通过强大的机器算法和大数据处理技术来实行数据价值挖掘。(4)处理速度快:处理速度是大数据区分与传统数据分析最显著的特征,如在电子商务背景下,不管其采用批处理还是流处理方式,其衡量的是用户“交互点”,如网站响应速度、订单完成速度、产品和服务的交付速度等。假设交互点是一个黑盒子,一边吸入数据,经过黑盒子处理后,在另一边流出价值,那处理速度指的是吸入、处理和产生价值的速度。在医疗信息服务中可能包含大量在线或实时数据分析处理的需求<sup>[2]</sup>。例如:临床决策支持中的诊断和用药建议、流行病学分析报表生成、健康指标预警等都需要更快的处理生成速度。因此,有大数据定义指利用常用软件工具捕获、管理和处理数据所耗时超过可容忍时间的数据集合。

## 2 大数据的产生及医疗卫生大数据来源

当今人类产生的数据量远超过以前人类历史任何时代所

产生的总和。数据的产生及来源大致经历了 3 个阶段:运营式系统阶段、用户原创内容阶段和感知式系统阶段<sup>[3]</sup>。在运营式系统阶段,数据库的出现使得数据管理的复杂程度大大降低,该阶段的数据往往伴随一定的运营活动而产生,并记录在数据库中,如在医疗服务过程中,对就医患者的信息登记和记录保存,该类数据的产生往往是被动的。在用户原创内容阶段,互联网及 Web2.0 时代的诞生促使人类社会数据量再次飞跃,如社交网络、专业网站、信息平台的建成后,人们愿意主动地在网上发布思想见解、经验和经历等,医学中的专业数据网站、医疗信息平台可以收集大量的数据信息,如医疗事故信息的上报、统计;医生专业知识的交流和分享;医学科研数据的网络化和搜索等,这个阶段数据产生方式往往是主动的。在感知式系统阶段,系统的广泛使用导致了数据信息的爆发,促进了大数据的产生<sup>[4]</sup>。这个阶段人们多使用微小的带有处理功能的传感器,并通过这些设备来对整个社会活动的运转进行监管,源源不断地产生新数据,如在医疗行业中的远程会诊、医学教学视频、移动医疗技术产生和传输的实时数据等等,这个阶段数据的产生方式多是自动的。因此,可以说大数据的产生伴随着电子信息和通信技术的发展,经历了被动、主动和自动 3 个阶段,它们构成了大数据的来源。

医疗卫生“大数据”的数据资源包括医疗服务的医院信息系统(HIS)、电子健康档案系统(EHRs)、实验室信息系统(LIS)、医学影像信息系统(PACS)、放射信息系统(RIS)的数据等,医院与医保的结算与费用数据,医学研究的学术、社会、政府数据,医院药物采购与使用监管数据,居民的行为与健康数据及政府的人口与公共卫生数据,构成了医疗卫生领域大数据的初期数据资源。而随着医改的不断深入,卫生统计和信息化建设得到了进一步加强,原卫生部“十二五”计划及卫生信息化建设工程规划中,初步确定了我国卫生信息化建设路线图,简称“3521 工程”,即建设国家级、省级和地市级三级卫生信息平台,加强公共卫生、医疗服务、新农合、基本药物制度、综合管理 5 项业务应用,建设健康档案和电子病历 2 个基础数据库和 1 个专用网络建设项目<sup>[5]</sup>。医疗和健康数据急剧扩容并成几何级增长,对其合理的挖掘利用,将会为广大患者、医务人员、科研人员及政府决策者提供更高效的服务和决策参考。

至 2003 年完成“人类基因组计划”以来,医学诊断模式发生了革命性的变化,“基因组后科学”的发展又提出了新的目标——精准医学,2011 年美国基因组学与生物医学界的智库发表了《迈向精准医学:建立生物医学与疾病新分类学的知识网络》<sup>[6]</sup>。随着分子生物学和高通量基因测序技术的发展,促进基因组学及基因组后的转录组学、蛋白质组、脂类组学、糖类组学、表观遗传学等多种“组学”进步,它们产生了海量的大数据,为社会精准医学的发展提供了丰富的数据源。随着现代信息技术的更新和世界全球一体化的推进,世界各国之间的信息网

\* 基金项目:“十二五”国家高技术发展计划(863 计划)资助项目(2014AA022304)。 <sup>△</sup> 通信作者, E-mail: 505069233@qq.com。

络平台建设得到了飞速发展,尤其是医疗卫生事业领域的信息交流更加密切,各种临床、科研、政府决策、分子生物学等医学信息的交流与共享,极大地丰富和整合了医学数据信息资源,为整个人类医学卫生事业的发展提供了良好的数据源保障。

### 3 大数据挖掘及医疗数据的利用

在大数据时代,大数据的价值体现在分析利用。在大数据分析过程中,需要从不同数据源获取数据,对数据进行预处理,并使用大量的挖掘模型对数据进行分析。早在 2012 年 3 月,美国奥巴马政府发布了“大数据研究和发展倡议”,投资 2 亿元以上启动计划,计划在科学研究、生物医学和环境科学等多领域突破性地挖掘利用大数据资源<sup>[7]</sup>。数据挖掘是指从大量数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的过程,它能够发现隐含在大规模数据中的知识,从而指导决策或提供参考,其主要涉及特征化、区分、关联或相关分析、分类、聚类、演变分析过程等。数据挖掘经历了单独算法、单个系统、单个机器的向量数据阶段;与数据库和数据仓库结合的,具有较高扩展性的能挖掘数据集的多维算法阶段;与预测模型相集成,支持 Web 数据、半结构化的数据,能挖掘网络环境下的分布式和高度异质的数据,并能有效地和操作型系统集成的一种网络化计算阶段;分布式数据挖掘,分布在多个节点上的网格计算的多种算法阶段;分布式并行数据挖掘与服务模式的云计算阶段。当前,云计算领域中的 Mapreduce, Hadoop 等高扩展性、高性能的并行计算编程模型、分布式海量数据框架及相关关键技术,是解决大数据挖掘任务的有效手段。目前,在 Mapreduce, Hadoop 编程模型基础上,已创新开发了多种适用于不同应用领域的能实现流处理、批处理或两种相容的处理模式的多种大数据挖掘处理工具。

目前大数据在医疗卫生领域有较广的应用,其涵盖了面向医生的临床辅助和科研,面向管理者的管理辅助决策、行业监管、绩效考核,面向居民的健康监测,面向药品研发的统计学分析、面向分子生物诊断的模型建设等<sup>[8-9]</sup>。如社交网络为许多慢性病患者提供了临床症状交流和诊治经验分享平台,医生借此可获得院外临床疗效统计数据。基于对人体基因的大数据分析,可以实现精准的个性化治疗。目前,国内大数据挖掘在医疗行业的应用研究方面得到了较快发展,王晔等<sup>[10]</sup>使用 HL7 CDA XML 作为描述电子病历的标准,采用关系型 XML 混合数据库提供索引和 XQuery 查询工具,同时使用了 BerkeleyDB 作为 Key-Value 存储的数据层,并架设了 Memcached 作为查询数据的缓存层,增强了整体系统的可用性,最终形成了一个标准、通用、高效的临床数据共享平台。陈曦等<sup>[11]</sup>采用了一种基于 Hadoop 的语义大数据分布式推理框架,并且设计了相应的基于属性链的原型推理系统来高效地挖掘中草药与西医基因之间的关联关系,发现了与肿瘤坏死因子(TNF)靶点分子相关联的 32 个中草药,这一数据挖掘技术有助于生物医学研究者从西医基因角度来分析和解释中医草药的潜在作用机制。近年来,大数据在分子生物学的发展中也发挥了重要作用,其在 DNA、RNA、蛋白质检测分析中数据的管理、储存、挖掘、处理和转化能力日益提升,它能准确、高效和快速地分析出个体的遗传、代谢信息后,结合上述医疗大数据中的其他数据资源,根据每位患者的个人化特征,精准地制订个性化治疗方案,以促进精准医学的发展。

### 4 精准医学的定义

精准医学指的是根据每例患者的个体化特征精准地制订个性化治疗方案。精准医学是现代医疗模式的革命和创新,其旨在实现对患者实施个体化诊断、临床医疗决策及临床预后评价,从而最大化地改善患者的生活质量。2015 年 1 月 20 日,

美国总统奥巴马在国情咨文演讲中提出了“精准医学”计划,它是继 2011 年美国基因组学与生物医学界的智库发表的《迈向精准医学:建立生物医学与疾病新分类学的知识网络》以来,更明确和详细的规划。2015 年 1 月 30 日,美国国立卫生研究院(NIH)和美国国家癌症研究所(NCI)主任哈罗德·瓦默斯(Harold Varmus)进一步解释了针对癌症的精准医学概念和计划内容<sup>[12]</sup>。精准医学是临床转化医学的组成部分,其将整合利用多种组后学技术、二代测序技术、基因组学、计算机生物学分析、医学信息学和临床信息学等多学科领域大数据资源<sup>[13]</sup>。随着基因测序行业的快速发展、生物医学分析的日渐成熟和生物大数据云计算技术日新月异的发展,将会从以前的“对症治疗”模式逐步转化为“对个体医疗”精准医疗模式,将针对每位患者独特的生物医学特征制订个性化的医疗方案<sup>[14]</sup>。这种精准医疗将以个人基因组信息为基础,结合蛋白质组学、代谢组等相关内环境信息,为患者量身设计出最佳治疗方案,以期达到治疗效益最大化和医疗资源配置最优化。

### 5 大数据挖掘促进精准医学发展

与精准医学相关的大项目已经有了较好的规划,奥巴马版的精准医学《报告》直接建议“百万美国人基因组计划”、“糖尿病代谢组计划”、“建立评估基因检测的新方法”、“癌症基因组计划”等。这些项目的完成离不开其对生物信息学大数据的获取和高效地挖掘。美国国家生物技术信息中心(NCBI)和欧洲生物情报协会(EBI)有着近 30 年历史的生物信息中心,他们在推动精准医学发展方面扮演着重要的角色。

精准医学涵盖了传统的流行病学、预防医学、临床诊断学和治疗学、康复医学及卫生经济学等学科信息资源,其支柱性技术领域包括医学信息学、医学分子生物学。目前,分子生物学中 DNA 测序已经具有能精确检测单个核苷酸、单细胞和单分子(或超微量)的分析技术,这将引领着未来体内检测技术的发展。高通量基因芯片和蛋白芯片的常规化应用使得各种不同疾病的基因分型成为可能,特定患者的分子疾病谱的绘制已逐渐成为临床诊疗中的重要工作部分<sup>[15]</sup>。然而上述检测技术产生的海量数据信息对大数据管理、储存、挖掘、处理和转化能力提出了更高的要求,目前随着云计算技术的产生和发展,能辅助临床医生、分子生物学家、统计学家、生物信息学家们整合所需的大数据,并挖掘分析以促进多学科领域的发展与更新<sup>[16]</sup>,为精准医学的发展提供强大的技术支持。精准医学数据挖掘框架,见图 1。如在临床中,3D 影像数据中心快速、高效的多种影像肿瘤跟踪软件通过跟踪对比不同时期的正电子发射计算机断层摄影(PET)/CT、单光子发射计算机断层(SPECT)/CT、磁共振成像(MRI)及 CT 检查结果,能轻而易举地监测、评估病情的发展和疗效,在时间轴上分割病灶、量化结果,为个体化治疗提供依据<sup>[17]</sup>。

为了促使精准医学的实施,美国科学院设计了一个“模型”,为基础研究发现和临床医学发展建立起一个共同的生物学信息数据库(网络),通过搜集每例患者的基因组学、表观基因组学、蛋白组学、信号传导学、临床症状体征及临床实验室检测数据,结合体内微生物学、外环境暴露学、社会学等资讯,建立个体信息港,通过大协作,建立疾病知识共享平台,在大数据的框架下,寻找疾病的分子基础及驱动因素,重新将疾病分类,实现精准的疾病分类及诊断,在此基础上,开展循证医学研究,对有相同病因、共同发病机制的患者亚群实现精准的评估、治疗及预防<sup>[6,18]</sup>。目前,在精准医学数据挖掘方面,人工神经网络技术、MetaLab、MetaCore 等多种生物信息学技术发挥了重要的挖掘、分析和预测功能<sup>[19-20]</sup>。

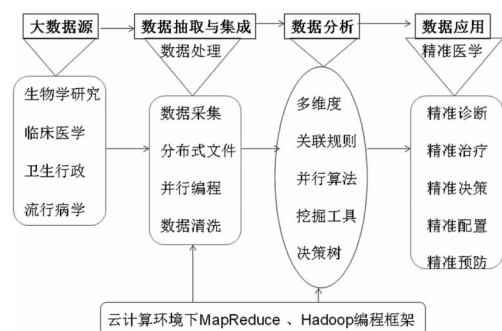


图 1 精准医学数据挖掘框架图

## 6 我国精准医学的发展动态

大数据时代的到来和精准医学时代雏形的出现,是人类社会发展历程的产物。我国在 2015 年 1 月 15 日,继国家卫生和计划生育委员会医政司发布第一批基因测序临床试点后,国家卫计委妇幼司也发布了第一批产前诊断试点单位,全国 31 个省市地区共有 109 家机构入选<sup>[21]</sup>,2015 年 3 月 11 日,科技部召开国家首次精准医学战略专家会议,并决定在 2030 年前,政府将在精准医疗领域投入 600 亿元。2015 年 3 月 27 日,国家卫生和计划生育委员会医政医管局又发布了第一批肿瘤诊断与治疗项目高通量基因测序技术临床试点单位名单,这体现了我国政府对精准医疗的快速反应和发展期待。充分发挥大数据挖掘技术在我国卫生诊治服务、生物医药研发、疾病防治研究、健康管理等多方面的应用,使其高效地为精准医疗体系建设服务,这是这一时代人应该为之努力的事业。

## 7 大数据挖掘在我国精准医学发展中面临的挑战

我国人口众多、疾病谱较复杂,虽然临床生物标本较为丰富,但中国目前的医疗制度和现状存在着严重的问题,医疗改革也在被热议和探索中,因此,亟待借助大数据挖掘技术探求出适合我国国情的精准医学及高效的医疗制度体系<sup>[22]</sup>。然而,我国的精准医学发展处于刚起步阶段,面临着诸多问题和挑战,亟须做好以下建设工作:打破常规思维模式,启用创新的思维方式;加强大数据管理能力,包括加强大数据源的质量监管和提升其数据处理的实时性;在复杂的网络环境下加强动态数据的利用和隐私保护;增强大数据管理的易用性,促进共享机制的建设;加强大数据挖掘技术的动态开发能力;推进适应我国精准医学发展要求的复合型人才队伍的建设,培养大批医学数据科学家等。总之,我国精准医学的发展道路艰辛而漫长,需要充分发挥大数据的挖掘能力,以加快精准医学及精准医疗体系的可持续发展。

## 参考文献

[1] 张肖,杨锦洲,王志勇. 基于大数据的医疗健康创新应用[J]. 电子科学技术,2015,2(2):190-194.  
[2] 王韬,陈祁. 大数据浪潮下的医疗数据应用模式与实践[J]. 中国

信息界:e 医疗,2014,7(8):57-60.

[3] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展,2013,50(1):146-169.  
[4] 张恒,张浩. 基于云计算的区域医疗主数据管理系统设计[J]. 中国数字医学,2013,8(9):18-20.  
[5] 高汉松,肖凌,许德玮,等. 基于云计算的医疗大数据挖掘平台[J]. 医学信息学杂志,2013,34(5):7-12.  
[6] 肖飞. 从循证医学到精准医学的思考[J]. 中华肾病研究电子杂志,2014,3(3):123-128.  
[7] Lancet T. Moving toward precision medicine[J]. Lancet,2011,378(9804):1678.  
[8] 周光华,辛英,张雅洁,等. 医疗卫生领域大数据应用探讨[J]. 中国卫生信息管理杂志,2013,10(4):296-300.  
[9] 曹原. 医疗大数据风起云涌[J]. 中国医院院长,2015,11(2):55-56.  
[10] 王晔,王崇,王晓玲,等. 多中心临床数据的快速检索系统[J]. 计算机研究与发展,2010,47(1):499-503.  
[11] 陈曦,陈华钧,顾佩琳,等. 一种基于 Hadoop 的语义大数据分布式推理框架[J]. 计算机研究与发展,2013,50(z2):103-113.  
[12] Collins FS, Varmus H. A new initiative on precision medicine[J]. N Engl J Med,2015,372(9):793-795.  
[13] 于军. “人类基因组计划”回顾与展望:从基因组生物学到精准医学[J]. 自然杂志,2013,35(5):326-331.  
[14] Blum M. Digital health center at UCSF clears the path for innovation. The new initiative aims to drive precision medicine forward. Interview by Rajiv Leventhal[J]. Healthc Inform,2013,30(6):44-46.  
[15] 蔡佳慧,张涛,宗文红. 医疗大数据面临的挑战及思考[J]. 中国卫生信息管理杂志,2013,10(4):292-295.  
[16] 许德泉,杨慧清. 大数据在医疗个性化服务中的应用[J]. 中国卫生信息管理杂志,2013,10(4):301-304.  
[17] 曹原. 3D 中心实现精准医疗[J]. 中国医院院长,2014,10(16):78-79.  
[18] 李静,顾江. 个体化医疗和大数据时代的机遇和挑战[J]. 医学与哲学,2014,35(1):5-10.  
[19] 潘湘高,彭楚武,管茶香. 人工神经网络在生物医学检测中的应用—生化测量数据处理方法的研究[J]. 中国生物医学工程学报,2007,26(2):161-165.  
[20] 郭霞,刘开江,哈丽丹·热依木,等. 基于 MetaCoreTM 分析维吾尔族宫颈癌患者潜在血浆标志物分析[J]. 中国肿瘤临床,2013,51(17):1020-1024.  
[21] 苏喧. 基因检测掀开精准医疗帷幕,带来医学领域颠覆性革命:中国医科大学第一医院肿瘤内科主任刘云鹏谈——最新基因检测整体态势和未来趋向[J]. 中国医药科学,2015,5(4):1-6.  
[22] 夏锋,韦邦福. 精准医疗的理念及其技术体系[J]. 医学与哲学:临床决策论坛版,2010,31(11):1-3.

(收稿日期:2015-06-28)

## • 检验科与实验室管理 •

# 贝克曼 DXI-800 化学发光免疫分析仪室内质量控制的实施和失控分析

刘海红

(连云港市第一人民医院检验科,江苏连云港 222002)

DOI:10.3969/j.issn.1673-4130.2015.23.067

文献标识码:B

文章编号:1673-4130(2015)23-3501-03

化学发光免疫分析(CLIA)是将高灵敏的化学发光技术与

高特异性的免疫反应结合起来,建立的化学发光免疫分析法。