

信息化技术在精准医疗发展中的应用研究^{*}

崔芳芳 陈昊天 王文超 赵 杰

(郑州大学第一附属医院/互联网医疗系统与应用国家工程实验室 郑州 450052)

〔摘要〕 阐述信息化技术在精准医疗数据采集、融合、存储、分析、临床决策系统建设、数据安全等方面的应用情况及作用,分析信息化技术支撑下的精准医疗发展面临的问题并提出建议,包括构建区域健康医疗数据管理平台,实现数据互联互通;创新大数据分析算法,研发智能分析工具等。

〔关键词〕 精准医疗;信息化技术;大数据;辅助决策

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.11.013

Study on the Application of Information Technology in the Development of Precision Medicine CUI Fangfang, CHEN Haotian, WANG Wenchao, ZHAO Jie, National Engineering Laboratory for Internet Medical Systems and Applications, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, China

〔Abstract〕 The paper expounds the applications and roles of information technology in precision medicine data collection, fusion, storage, analysis, clinical decision system construction, data security and other aspects, analyzes the problems faced by the development of precision medicine supported by information technology, and puts forward suggestions, including the building of regional medical and health data management platform to achieve data interconnection; innovation of big data analysis algorithms and development of intelligent analysis tools, etc.

〔Keywords〕 precision medicine; information technology; big data; assistant decision-making

1 引言

精准医疗又称个性化医疗或个体化医疗,即根

〔修回日期〕 2021-06-09

〔作者简介〕 崔芳芳,硕士,统计师,发表论文16篇;通讯作者:赵杰,博士,教授。

〔基金项目〕 国家重点研发计划“精准医学研究”重点专项“基于远程/移动医疗系统的精准医学联合构建、综合服务平台与业务系统开发及应用”(项目编号:2017YFC0909900);郑州市协同创新专项“基于高通量测序方法的血浆游离核酸检测关键技术”(项目编号:20XTZX08017)。

据个体基因特征、环境以及生活习惯进行疾病干预与治疗。随着测序技术的成熟和组学数据的积累,精准医疗研究在全球范围内逐渐上升到国家战略高度,美国、韩国、英国等分别部署了精准医疗战略规划,推动精准医疗研究及应用转化^[1]。我国于2015年提出设立精准医学重点研发计划,开展精准医疗大数据平台研发、罕见病队列研究、常见多发肿瘤队列研究等^[2]。随着精准医疗研究的不断深入,除基因测序等基础研究外,数据融合、分析平台建设、决策系统开发等方面也成为研究热点^[3]。海量的测序数据需要通过信息化手段进行整合、挖掘,构建精准医疗大数据平台和辅助决策系统,实现电子病历信息与组学数据的共同存储与融合,推动精准医疗研究成果的临床转化是未来精准医疗发

展重点。信息化技术已广泛应用于医疗领域,如医疗信息系统、数据平台、决策系统等^[4-5]。精准医疗作为医疗领域的创新性研究,部分成果已应用于临床决策,但研究成果的临床转化仍存在较大局限性。未来研究重点包括基因测序结果解读、生物样本库搭建、成果转化与临床决策等,均需信息化技术支持。本文分析有关信息化技术在精准医疗数据采集、存储、分析与处理、大数据平台与决策系统建设、数据安全等领域的应用情况,为精准医疗数据分析、决策平台构建提供参考。

2 信息化技术在精准医疗中的应用

2.1 数据处理

2.1.1 数据采集 精准医疗研究的基础,涉及数据类型包括组学、个体表型及环境数据。组学数据包括基因组、转录组、蛋白组、代谢组数据等,主要采用高通量测序方法获得原始数据,进而通过提取、峰对齐、去噪等处理将原始数据转换为可分析的数据类型。表型数据采集可利用智能监测设备。通过佩戴智能眼镜、手表、头盔等可穿戴设备监测个人生理参数、生活习惯等数据,利用手机等移动通信设备将数据上传至云端服务器^[6-8]。此外有学者研发智能健康监测系统采集健康数据,系统由智能监测终端、医疗数据处理服务器、移动客户端3部分组成,配备全球定位系统(Global Positioning System, GPS)以及3G通信模块,实现对人体体温、心率、血氧、定位数据的采集并上传服务器,在移动客户端实现数据共享^[9]。为提高动态测量准确性,有学者研发柔性电子技术,实现对生理指标、生理电信号和生物化学分子等各类人体健康指标的监测^[10]。环境数据可通过环境自动监测系统采集,如大气自动监测、水质自动监测系统等,自动采集、传输、分析环境监测结果^[11]。信息化技术的发展推动数据采集转向自动化,提高采集效率,有效缓解人工采集导致的耗时、耗力、错误率高等现象。

2.1.2 数据融合 即将临床诊疗、组学、健康数据等多模态数据按照统一标准及处理流程,对

结构化数据进行抽取入库,对非结构化数据进行结构化改造,将身份证ID或住院号等作为统一标识,进行各类数据集成融合,实现数据共享存储。大数据融合处理主要采用数据仓库技术,将数据从来源端经过抽取、转换、加载(Extract - Transform - Load, ETL)至目的端,常用处理工具有Kettle、DataX、Apache Camel等,用于数据抽取、传输、同步、运行等。有研究基于理论视角,结合精准医疗数据特点,从层次维、时间维和种类维3个方面构建精准医疗数据融合标准框架体系^[12]。另有学者开展精准医疗数据融合实践,从ClinVar和COSMIC突变数据库中获取基因突变,根据数据特征建立数据标准,进行突变数据分类与融合^[13]。有研究提出精准医疗大数据平台构建方案,采用分布式数据库进行不同类型数据加工、关联、融合,支持数据分布式计算^[14]。综合看来,在信息化技术支持下已有研究开展精准医疗数据融合理论与实践工作,但有关精准医疗数据集成平台建设与应用的报道较少,面向精准医疗的多源异构数据融合仍是研究关键。

2.1.3 数据存储与检索 基因测序研究在国内外精准医疗研究中备受重视,如英国10万基因组计划、冰岛35万基因组计划、中国10万人基因组计划等。越来越多的基因测序研究产生海量测序数据,建立有效的存储数据库和平台是精准医疗数据分析的重要前提。组学数据具有数据量大、更新快、类型复杂的特点,数据上传、存储、查询提取与分析等均需信息化技术支持。目前全球较权威的3个组学数据库为美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)遗传序列数据库GenBank,日本DNA数据库(DNA Data Bank of Japan, DDBJ),欧洲分子生物学实验室数据库(European Molecular Biology Laboratory, EMBL)^[15-17],3个数据库汇集所有公开的核酸以及蛋白质序列,主要通过BankIt、Sequin等软件进行数据提交,利用Nucleotide进行数据查询,通过局部序列排比检索基本工具(Basic Local Alignment Search Tool, BLAST)进行数据相似性搜索。由蛋白质信息资源(Protein Information Resource, PIR),

慕尼黑蛋白质序列信息中心 (Munich Information Centre for Protein Sequences, MIPS) 和日本国际蛋白质序列数据库 (JIPID) 共同维护的国际上最大的公共蛋白质序列数据库 (PIR - PSD), 是支持基因组学、蛋白质组学、系统生物学检索和科学研究的综合公共生物信息学资源。PIR - PSD 提供 3 种类型的检索服务: 基于文本的交互式查询、标准序列相似性搜索以及结合序列相似性、注释信息和蛋白质家族信息的高级搜索。此外由美国国家癌症研究所 (National Cancer Institute, NCI) 和美国国家人类基因组研究所 (National Human Genome Research Institute, NHGRI) 建立的肿瘤基因组数据库 (The Cancer Genome Atlas, TCGA) 也是常用的开放共享数据库。在国内, 北京基因组研究所生命与健康大数据中心 (BIG Data Center, BIGD) 建立的多层次生物组学数据资源系统, 包括原始组学数据归档库 (Genome Sequence Archive, GSA), 基因组数据库 (Genome Warehouse, GWH) 等^[18]。上述数据库应用信息技术分析生物学信息, 通过计算机运算法则探索突变基因, 采用 Oracle、Sybase 等管理软件进行数据库维护。现有数据库多采用分布式计算, 将不同来源数据整合在同一平台。有研究提出利用基于分层思想的软件层次架构构建基于 Web 的代谢组学数据库^[19]。部分二次数据库, 如基因调控转录因子数据库 TransFac、同源蛋白结构数据库 Blocks 等, 可基于 Web 浏览器、利用超文本语言 HTML 和 Java 程序编写界面, 使界面更加友好^[20]。精准医疗研究需要结合临床诊疗和健康数据, 利用信息化技术进行电子病历信息整合, 形成可利用基因信息的电子健康档案以辅助精准医疗研究。我国精准医疗重点研发项目中的重点研究方向涉及精准医疗大数据资源整合、存储、利用与共享平台建设, 信息化技术是精准医疗信息整合的关键。

2.2 数据分析

2.2.1 总体情况 精准医疗数据分析的目的是整合医院临床诊疗、生物组学、人群健康、环境数据以及文献资料等多维度数据, 借助医疗自然语言处理、机器学习、数据可视化和分析等技术和工具,

构建基于大数据和人工智能的学习型医疗体系, 实现精准医疗的应用转化。基因测序技术从第 1 代发展到第 4 代, 测序通量爆炸性增长, 产生海量测序数据, 仅用传统统计方法已无法满足测序数据分析需求, Hadoop 技术、云计算等成为测序数据分析的主要方法。然而精准医疗领域研究者多为医学、遗传、生物等专业, 不具有计算机专业背景, 测序数据处理成为研究难题, 研发自动化数据处理系统成为解决数据分析难题的重要手段。

2.2.2 国外相关研究 Mohr C、Friedrich A 和 Wojnar D 等^[21] 利用关系数据库、数据存储、数据转换、数据模型技术研发数据驱动的生物医学研究平台, 可提供流程化的实验设计、数据管理、质控与自动分析、结果可视化展示。美国有机构研发云端 DNA 数据库平台, 采用 Google 云服务将数据存储在云端, 进行测序数据分析。此外有学者研发 DistMap、FX、CloudAligner 等序列数据分析工具^[22-24]。DistMap 主要对短读长序列数据进行比对, 利用 Hadoop 集群实现分布式数据分析, 大幅加速序列比对速度; FX 是基于 Hadoop 系统的 RNA 序列数据分析软件; CloudAligner 主要用于长序列数据比对, 能够对 DNA 基因组和 RNA 转录组序列进行分析。

2.2.3 国内相关研究 黄芝准^[25] 基于 Hadoop 系统设计高通量测序数据自动化并行处理系统, 通过 MapReduce 并行运算框架对测序数据进行分割、对比、信息查询, 输出突变基因信息; 林晶晶^[26] 基于 Hadoop 技术开发人类全基因组重测序数据处理系统。华大基因研发了基因组数据分析云计算平台, 支持数据云端存储、分析、展示和交付, 为不同规模研究人员和机构提供基因组学数据分析。此外有机构推出精准医疗大数据平台, 整合医院异构临床、生物组学、病理检查、影像检查、实验室检测、随访数据以及各种知识库、指南、文献等多维度数据, 提供强大的机器学习、数据可视化工具和探索工具, 构建基于人工智能技术的学习型医疗体系。

2.3 临床辅助决策系统构建

精准医疗研究的最终目的是形成精准防治方案和临床决策系统, 辅助临床疾病诊治和用药, 显著

提升人群健康水平。临床决策系统和精准用药系统的构建离不开信息化技术支持，常用信息化技术有云计算、深度分析、神经网络等。目前国内外研发精准医疗决策系统，如范德比尔特－英格拉姆癌症中心研发的 My Cancer Genome 系统、浙江大学药

院研发的精准用药搜索系统等，均采用信息化技术为癌症患者提供精准治疗措施和用药方案。总结国内外精准医疗决策系统功能及其采用的信息化技术，见表 1^[27-31]。

表 1 国内外精准医疗临床决策和用药支持系统

系统名称	研究机构	系统功能	相关技术
精准用药搜索系统	浙江大学药学院	根据癌症患者的转录组特征，为医生和患者快速提供用药参考	数据库映射、组合逻辑拆解
IBM 电脑医生 Watson	IBM 公司、安德森癌症中心	肿瘤诊断，确定患者基因的驱动突变及作用的药物靶标	标化、整合患者病历、实验数据，深度分析和挖掘
My Cancer Genome	范德比尔特－英格拉姆癌症中心	整合基因突变信息和电子健康记录，提供最新的影响不同癌症的基因突变和针对突变的治疗措施，为肿瘤科医生提供临床决策支持	文本挖掘、人工智能、风险预测
G－DOC Plus	乔治城大学医学中心，隆巴尔迪综合癌症中心	整合 10 000 例以上患者诊疗数据，以及实验室检查、病理和影像数据、基因测序数据等，进行肿瘤和非肿瘤疾病的生物标志物探索、多组学分析、体细胞突变研究等，为临床诊疗提供决策支持	可视化技术、Groovy and Grails 应用程序、Java 虚拟机、JBossMQ、EC2 实例、云计算等

2.4 数据安全保障

精准医疗研究的基础是海量基因数据和患者健康信息。在采集患者信息时涉及遗传、生活习惯、病史等个人隐私，保护信息安全、不泄露患者可识别信息是精准医疗研究的重要问题。保障数据安全除建立完善的数据管理和信息访问制度外，更重要的是通过信息化技术手段保障数据安全。例如建立数据库访问控制机制、数据加密、建立数据备份系统、通过防火墙防止恶意侵入、数据传输加密、审计跟踪与节点验证等。魏明月、刘逸敏和周骏群等^[32]在建设肝肿瘤标本信息库时，利用基于角色权限动态转换的隐私数据访问控制方法实现动态授权并限时访问隐私数据。林炜炜^[33]在阿尔茨海默病精准医疗大数据研究平台建设中，通过数据库备份机制防范数据损失和网络攻击，利用 X. 509 数字证书技术对数据传输层加密，同时在中心端高效解密，保障数据传输安全。精准医疗大数据存储分析、决策系统构建都需在保障数据安全的基础上进行，利用信息化技术强化数据安全措施也是开展精准医疗临床应用的重要内容。

3 信息化技术支撑下的精准医疗发展面临的问题

3.1 数据采集融合困难

目前国内外在信息化与精准医疗相结合的研究领域已经取得较大成果，但精准医疗研究仍存在较多问题，尤其在数据采集融合方面，如数据孤岛、数据缺失等。在临床数据采集方面，数据缺失、数据壁垒现象较为常见，医院信息系统未实现互联互通，不同研究数据独立管理等，导致临床数据融合存在问题。不同医疗机构间检验检查设备仪器不同，导致数据标准不同，数据内部存在异质性。在数据融合方面，存在多组学数据与临床数据无法真正有效结合、临床数据未实现结构化处理、医院未建立标准生物数据库等问题，严重阻碍组学数据与表型数据结合，进而导致数据完整性较差。

3.2 数据分析利用率不高

随着大数据技术快速发展，算法工具和数据分析技术不断创新，促进多模态医疗数据发展。但在精准医疗大数据分析方面融合算法研究尚不成熟，

缺乏同时处理大规模组学和临床数据的算法,导致数据分析利用效率不高。针对多模态数据不完整、数据处理实时性等情况,需探索更为有效的数据处理方法。精准医疗大数据分析属于医工交叉领域,需要临床、计算机、生物信息等多学科专业技术人员协作完成,目前医工交叉领域人才短缺,跨学科协作机制不清晰,影响数据分析与利用。

3.3 成果转化应用有限

精准医疗研究的最终目的是形成更为科学有效的重大疾病防治与用药方案,提升临床诊疗水平,规范临床治疗路径。但目前精准医疗研究成果转化有限,其应用主要体现在肿瘤治疗方面,在其他重大疾病治疗方面的作用有待进一步研究。同时由于数据融合与分析难度较高,缺乏真正能够辅助临床诊疗的精准医疗决策系统,限制了其在临床的快速转化。针对上述问题亟需探索新的数据算法和建模方法,充分挖掘数据价值,研发新型精准诊疗方案并转化为辅助决策系统,促使精准医疗由研究向临床应用转化。

4 精准医疗发展建议

4.1 构建区域健康医疗数据管理平台,实现数据互联互通

现阶段我国区域内医疗机构信息系统相互独立,数据孤岛现象普遍,数据共享融合困难,针对该问题应积极探索数据共享技术,在全国范围内筛选试点地区,建立区域示范健康医疗数据综合管理平台,整合统一业务系统,制定数据标准,将医疗机构、公共卫生机构等业务系统数据进行汇聚,从政府层面建立数据共享机制和数据管理权限,实现区域范围内跨机构、跨系统健康医疗数据整合管理、实时查询、统一监管,提供集约式、一站式数据服务。同时加强数据质量控制,对上传至平台的数据进行信息校对,解决现有数据质量不高、缺失严重的问题。

4.2 创新大数据分析算法,研发智能分析工具

精准医疗研究的基础是利用大数据分析技术,

在分子水平获取与疾病相关的知识,建立基因型与表型之间的关联,从分子水平制定疾病诊疗方案。因此需不断研究高效的数据分析技术,深度挖掘数据,探索隐藏在海量数据中的医学信息,实现辅助疾病诊疗的目的。一方面,针对精准医疗测序数据,发展医学和生物学信息解读技术,提升测序信息解读能力;另一方面,针对精准医疗研究多模态数据特点,研发具有针对性、运算效率更高、结果更准确的大数据并行分析方法,充分挖掘数据信息。此外在数据分析算法基础上,开发自动化分析工具,为科研人员和医生等非信息化背景专业人员提供更为便捷的数据分析途径。

4.3 加强产学研合作,推进精准医疗技术创新与成果转化

目前精准医疗研究主要集中在测序技术、疾病分子分型、靶点挖掘等基础研究方面,研究成果的临床转化应用较为局限。为推动精准医疗快速发展与技术转化,应建立涵盖科研机构、医疗机构、医药企业的精准医疗研究与转化合作组织与产业链,在基础研究、技术攻关、成果转化、临床诊疗、产业发展、人才培养等方面建立跨机构合作关系,根据临床诊疗需求开展科学研究与技术创新,突破重大疾病诊疗难题,依托医药企业等转化为直接服务临床的诊疗产品,在临床进行应用与推广,辅助临床诊疗和用药,推动精准医疗成为未来医疗的主要发展模式。

5 结语

精准医疗已成为我国未来医疗模式发展的新方向,信息技术在推动精准医疗发展与应用方面有重要价值。本研究从数据采集融合、数据分析、智能辅助决策、信息安全等方面综合分析信息技术在精准医疗现有研究中发挥的支撑作用,充分说明精准医疗的发展离不开信息技术的支持。同时针对精准医疗发展中存在的问题提出信息化方面的解决思路和建议,为加速精准医疗发展提供参考。

参考文献

- 1 徐鹏辉. 美国启动精准医疗计划 [J]. 世界复合医学,

- 2015 (1): 44 – 46.
- 2 谢俊祥, 张琳. 精准医疗发展现状及趋势 [J]. 中国医疗器械信息, 2016, 22 (11): 5 – 10.
- 3 Aronson S J, Rehm H L. Building the Foundation for Genomics in Precision Medicine [J]. *Nature*, 2015, 526 (7573): 336 – 342.
- 4 张洪磊. 基于 Hadoop 的医院数据中心系统设计与实现 [D]. 杭州: 浙江大学, 2014.
- 5 Rajeevan N, Niehoff K M, Charpentier P, et al. Utilizing Patient Data from the Veterans Administration Electronic Health Record to Support Web – based Clinical Decision Support: Informatics Challenges and Issues from Three Clinical Domains [J]. *BMC Med Inform Decis Mak*, 2017, 17 (1): 111.
- 6 Son D, Lee J, Qiao S, et al. Multifunctional Wearable Devices for Diagnosis and Therapy of Movement Disorders [J]. *Nat Nanotechnol*, 2014, 9 (5): 397 – 404.
- 7 Bayo – Monton J L, Martinez – Millana A, Han W, et al. Wearable Sensors Integrated with Internet of Things for Advancing eHealth Care [J]. *Sensors (Basel)*, 2018, 18 (6): E1851.
- 8 人民网. 谷歌智能隐形眼镜可测血糖, 技术目前处初期阶段 [EB/OL]. [2021 – 04 – 28]. <http://it.people.com.cn/n/2014/0122/c1009-24191436.html>.
- 9 周云如. 智能健康监测终端设计与实现 [D]. 西安: 西安科技大学, 2016.
- 10 赵一聪, 徐可欣, 黄显. 精准医疗中的柔性电子技术 [J]. *科技导报*, 2017, 35 (23): 76 – 81.
- 11 欧阳俊, 张子凡, 王淮滨. 智能网络环境监控数据采集传输仪的设计 [J]. *现代仪器*, 2008, 14 (6): 10 – 12, 17.
- 12 翟运开, 路薇, 张瑞霞, 等. 多维集成视角下精准医疗数据融合标准体系构建 [J]. *中国卫生资源*, 2020, 23 (1): 23 – 27.
- 13 吴萌, 李姣, 康宏宇, 等. 面向精准医学的基因突变数据分类与融合研究 [J]. *中华医学图书情报杂志*, 2018, 27 (11): 16 – 22.
- 14 李伟, 张麟, 李霞, 等. 精准医学大数据平台关键技术研究 [J]. *医疗卫生装备*, 2017, 38 (8): 41 – 46.
- 15 饶冬梅. NCBI 数据库及其资源的获取 [J]. *科技视界*, 2013 (7): 53 – 54.
- 16 Kodama Y, Mashima J, Kosuge T, et al. The DDBJ Japanese Genotype – phenotype Archive for Genetic and Phenotypic Human Data [J]. *Nucleic Acids Res*, 2015, 43 (Database Issue): D18 – D22.
- 17 Li W, Cowley A, Uludag M, et al. The EMBL – EBI Bioinformatics Web and Programmatic Tools Framework [J]. *Nucleic Acids Res*, 2015, 43 (W1): W580 – W584.
- 18 余惠敏. 中国 BIGD 跻身国际生物组学数据库 [N]. *经济日报*, 2017 – 02 – 13 (15).
- 19 徐肖肖. 基于 Web 的代谢组学数据库软件设计与实现 [D]. 济南: 济南大学, 2015.
- 20 张敏辉, 高晓玲. 基于生物信息学数据库的研究与应用 [J]. *中国科技信息*, 2010 (22): 44 – 45.
- 21 Mohr C, Friedrich A, Wojnar D, et al. qPortal: a platform for data – driven biomedical research [J]. *PLoS One*, 2018, 13 (1): e0191603.
- 22 Pandey R V, Schlötterer C. DistMap: a toolkit for distributed short read mapping on a Hadoop cluster [J]. *PLoS One*, 2013, 8 (8): e72614.
- 23 Hong D, Rhie A, Park S S, et al. FX: an RNA – Seq analysis tool on the cloud [J]. *Bioinformatics*, 2012, 28 (5): 721 – 723.
- 24 Nguyen T, Shi W, Ruden D. CloudAligner: a fast and full – featured MapReduce based tool for sequence mapping [J]. *BMC Res Notes*, 2011 (4): 171.
- 25 黄芝准. 组学大数据环境下的基因信息并行处理与分析方法研究 [D]. 合肥: 中国科学技术大学, 2017.
- 26 林晶晶. 基于 Hadoop 的人类全基因组重测序数据处理技术的研究及实现 [D]. 大庆: 东北石油大学, 2015.
- 27 海玲, 陈新. 基于病例转录组特征的精准用药搜索系统 [J]. *基因组学与应用生物学*, 2018, 37 (2): 637 – 642.
- 28 American Association for Cancer. Oncologists Partner with Watson on Genomics [J]. *Cancer Discov*, 2015, 5 (8): 788.
- 29 Doyle – Lindrud S. Watson Will See You Now: a supercomputer to help clinicians make informed treatment decisions [J]. *Clin J Oncol Nurs*, 2015, 19 (1): 31 – 32.
- 30 Carney P H. Information Technology and Precision Medicine [J]. *Semin Oncol Nurs*, 2014, 30 (2): 124 – 129.
- 31 Bhuvaneshwar K, Belouali A, Singh V, et al. G – DOC Plus – an integrative bioinformatics platform for precision medicine [J]. *BMC Bioinformatics*, 2016, 17 (1): 193.
- 32 魏明月, 刘逸敏, 周骏群, 等. 一种基于角色权限动态转换的隐私数据访问控制模型 [J]. *中国数字医学*, 2010, 5 (9): 11 – 15.
- 33 林炜炜. 阿尔茨海默病精准医学大数据研究平台的建设与应用 [D]. 济南: 山东大学, 2016.