健康医疗大数据在精准医疗领域的应用与挑战*

高景宏 李明原 王 琳 翟运开

（1 郑州大学第一附属医院 郑州 450052 2 互联网医疗系统与应用国家工程实验室 郑州 450052）

（1 互联网医疗系统与应用国家工程实验室 郑州 450052 2 郑州大学管理工程学院 郑州 450001）

〔摘要〕 介绍精准医疗概念与业务流程，阐述健康医疗大数据在精准医疗领域的应用及挑战并对未来提出展望。研究结果对提高健康医疗大数据利用效率、启发未来精准医疗研究方向、推进精准医疗在重点疾病防治中的应用具有一定意义。

〔关键词〕 健康医疗大数据；精准医疗；应用；挑战

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.05.003

Applications and Challenges of Health and Medical Big Data in the Field of Precision Medicine

GAO Jinghong, LI Mingyuan, WANG Lin, 1The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, 2National Engineering Laboratory for Internet Medical Systems and Applications, Zhengzhou 450052, China; ZHAI Yunkai, 1National Engineering Laboratory for Internet Medical Systems and Applications, Zhengzhou 450052, 2School of Management Engineering, Zhengzhou University, Zhengzhou 450001, China

〔**Abstract**〕 The paper introduces the concept and business process of precision medicine, expounds the application and challenges of health and medical big data in the field of precision medicine, and puts forward prospects. The findings of the study are of certain significance for improving the utilization efficiency of health and medical big data, enlightening the future research direction of precision medicine, and promoting the application of precision medicine in the prevention and treatment of priority diseases.

〔**Keywords**〕 health and medical big data; precision medicine; application; challenge

1 引言

目前我国公众健康和疾病负担形势严峻。报告显示我国现有高血压患者 2.6 亿人，每年心血管疾病死亡人数达 300 万，糖尿病患者数量超过 1 亿，每年因癌症死亡人数达 220 万例[1]。在此背景下可通过对健康医疗大数据的集成处理与深入挖掘，有效促进精准医疗在疾病防治中的应用，进而优化医疗资源，减少无效和过度医疗，提升医疗服务效率与质量，最终提高大众健康水平。健康医疗大数据是精准医疗深入应用的关

〔修回日期〕 2021－08－01

〔作者简介〕 高景宏，博士，助理研究员，发表论文 41 篇，参编著作 2 部。

〔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/移动医疗网络的精准医疗综合服务示范体系建设与推广”（项目编号：2017YFC0909900）；河南省自然科学基金青年项目“基于多源健康医疗数据的郑州市雾霾污染健康风险评估与机制研究”（项目编号：202300410409）；国家超级计算郑州中心创新生态系统建设科技专项“基于多模态数据的肺癌智能诊疗关键技术研究与示范应用”（项目编号：201400210400）；河南省高校科技创新团队支持计划“医疗大数据分析与临床应用”（项目编号：20IRTSTHN028）。

键，本研究对健康医疗大数据在精准医疗领域的应用、挑战及未来研究方向等进行探讨，以期为夯实精准医学研究基础、推进精准医疗在疾病防治中的应用、提高疾病诊断与治疗效率等提供参考。

2 精准医疗概念与业务流程

2.1 概念

精准医疗是应用基因检测、现代遗传、分子影像、组学、大数据等技术，根据患者临床诊疗、生物信息、生活环境与习惯等相关数据，实现精准疾病分类与诊断，筛选对疾病进行干预和治疗的最佳靶标与方法，为临床实践提供科学依据，为患者定制个性化的疾病治疗和预防方案，使患者获得最适宜的治疗效果和最低副作用的一种医疗模式[1-2]。精准医疗可以阐明疾病发生发展机制，解答疾病转归的本质问题；精确定位生物标志物，探索建立早期诊断方法，争取疾病治疗有效时机；通过分子分型和分期进行分子诊断，为个性化诊断、治疗和预后健康管理等提供科学依据。基于对健康医疗大数据的处理分析进行疾病综合防治方案的探索应用，见图 1[1,3-4]。

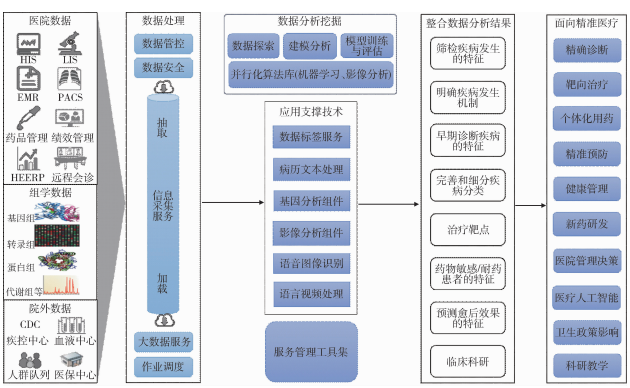

图 1 精准医疗图谱

2.2 业务流程（图 2）

2.2.1 形成精准的诊断结果 基于健康医疗大数据构建面向精准医疗服务的专病数据仓库，采用大数据分析和生物信息技术，深入挖掘患者疾病分型、病变靶点、易感基因、生物标志物等并生成可视化分析结果报告，结合临床医生和专家解读形成精准的诊断结果。

2.2.2 患者参与制定、选择治疗方案 根据诊断结果明确患者疾病分型，由临床医生、生物信息专家和患者一同参与治疗方案的制定与选择。在此过程中通过对治疗效果的实时评价与反馈及时调整、完善治疗方案，达到以患者为中心的最佳治疗效果。

2.2.3 识别用药靶点 通过对患者健康医疗数据进行分析，识别用药靶点，明确患者易感或病变基因、疾病症状与药物的关系，指导个性化用药并对药物治疗效果进行评价。

2.2.4 制定精准健康管理方案 基于对患者个体特征与需求的分析，制定贯穿患者整个诊疗过程的精准健康管理方案，如精准护理、康复管理、健康教育与促进等，形成以患者具体情况与需求为导向的全流程健康管理。

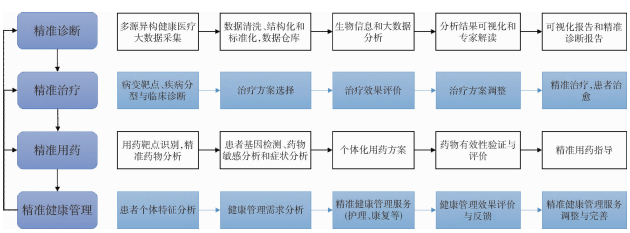

图 2 精准医疗服务业务流程

3 健康医疗大数据在精准医疗领域的应用

3.1 总体应用情况

健康医疗大数据具有全样本、深入关联、注重相关性等优势，可提升医务人员、科研工作者、卫生决策者和社会公众等应对疾病的洞察力和统筹规划能力，进而优化卫生资源配置和医疗服务流程，提高服务质量，控制医疗风险，降低诊治成本，全面提升疾病防治能力和医疗服务水平[5-6]。健康医疗大数据及其处理分析是开展各类精准医疗服务的前提，是进行精准诊断、精准治疗、个性化用药及精准健康管理等必不可少的环节。

3.2 精准诊断

首先，建立知识来源质量评估机制和精准医疗知识库，将有关专病的组学、临床、健康、环境等知识通过图理论关联，构建面向精准医疗的知识图谱。基于 Hadoop 和 Spark 的分布式文件和并行计算系统，研发针对精准医疗的文本处理算法，建立基于统计与基于规则相结合的精准医疗辅助专家决策系统[3,7]。其次，采集患者临床诊疗、生物样本、生活习惯和环境、组学信息等数据并对这些数据做结构化、标准化清洗与融合，构建专题数据集市。最后，基于患者多源健康医疗数据，利用大数据分析和生物信息技术对患者信息进行集成分析、可视化呈现，在临床医生和生物信息专家的共同参与下形成针对患者具体病情与症状的精准临床诊断，辅助临床医生研判患者疾病发生、演变过程及所处阶段[8]。

3.3 精准治疗

基于精准的疾病分类和诊断，结合患者临床诊疗、实验室检查、组学检测等信息，通过大数据分析得出针对患者具体情况的最佳诊疗方案。针对清洗与融合后的患者多源信息，利用组学、生物信息和大数据分析技术进行病变靶点、生物标志物、敏感生理生化反应指标等的分析、识别、验证与应用，尤其是针对高血压、脑卒中、心梗、肺癌等典型高发、危害严重的慢性病和常见肿瘤[8]。通过对生物样本、临床诊疗、组学等信息的深入挖掘，结合精准医疗专题知识库和专病知识图谱可以明确患者疾病病因、精准定位病变靶点，为患者提供最佳的个性化治疗方案，实现包括数据分析及可视化、治疗方案、个体化用药等在内的一体化精准医疗服务。

3.4 精准用药

传统临床用药根据患者症状体征、生化生理检查和影像资料对具有相同或相似临床症状的患者采用相同药物治疗。但实际情况是人体的药物敏感性和药物作用效果与个体基因、遗传、生活环境等密切相关，不同患者对同一药物的敏感性可能不同。精准医疗将传统的“对症下药”转变成“因人施药”，依据患者自身基因遗传特点、生存环境和生活习惯等进行个性化用药，是以基因测序技术为基础、大数据分析技术为手段的新型用药模式。具体来说精准用药是对患者临床诊疗、基因及个人体质特性等信息进行大数据分析，集成最优用药方案，为患者提供最切合自身情况的用药指导。基于对健康医疗大数据分析可明确不同患者对药物的敏感性差异和作用靶点，探明用药过程中可能出现的疗效、机体反应、毒副作用等，从而达到最正确时间节点、最佳用药剂量、最小不良反应的精准用药目标[1,9]。

3.5 精准健康管理

精准健康管理根据个体基因遗传背景，结合个体健康状况、患病情况、生活习惯和环境等信息进行系统、全面、持续的监测与评估，经过大数据集成处理与分析向个体提供健康咨询、生活方式与行为习惯指导、危险因素识别与干预、疾病护理与康复等个性化健康管理，是精准医疗的终极目标[1,10]。健康医疗数据的不断积累，尤其是组学数据的规范化积累与创新应用，为利用多源健康医疗大数据构建贯穿个体整个生命周期的预防、护理、康复、医疗保健等精准健康管理体系提供信息基础与技术支持。通过对健康医疗大数据的深入挖掘，可推动面向患者整个健康与疾病过程的健康管理更加精准、高效[1]。

4 精准医疗领域大数据应用面临的挑战

4.1 总体情况

精准医疗是数据驱动的医疗服务模式，为挖掘健康医疗大数据中蕴含的有效信息以进行精准防治，需要对医疗数据进行深入分析与应用。随着人工智能、云存储、云计算等技术的发展，基于多源健康医疗信息的大数据集成分析变得更加高效、稳健，使临床医生能够精准地针对具体患者制定个性化诊疗方案，从而提高医疗服务效率和质量[1,6]。但是大数据在精准医疗领域的应用涉及数据采集、

清洗、分析、平台支撑、质量控制、数据治理等环节，均面临不同程度挑战，阻碍精准医疗进一步发展与应用。

4.2 数据采集

数据采集是根据研究目标抽象出的、在数据分析与应用中所需要的表征信息，通过多种方式从数据产生环境获取原始数据并进行预处理的一系列技术，是大数据分析与精准医疗应用的基础，为后续数据处理提供原始数据集[11]。在精准医疗领域，健康医疗大数据来源广泛，如何从中针对性地采集所需信息是首先需要考虑的问题，而传统数据采集手段缺乏相关技术储备。目前精准医疗领域大数据采集的内容和质量评价体系互不统一，同一类型数据往往存在多种不同采集方式，造成大数据样本之间存在不同程度的异质性[1]。例如健康医疗数据包括结构化（表单、列表）、半结构化（实验室检测报告、护理日志、体检报告）和非结构化（电子病历文本、医学影像、音视频）等类型，这些来源不同的数据以多种形式并存，分别需要专门技术进行采集。如何对这些数据采集技术进行集成整合，从而在兼容多种数据传输协议、接口方案的前提下进行高效的采集与传输服务，成为亟待解决的问题[11]。数据采集过程涉及信息安全和患者隐私，这不仅是医学伦理问题，还是数据采集技术层面问题，目前在健康医疗数据采集过程中尚无成熟手段对此予以保障。

4.3 数据清洗

数据清洗是对采集的原始数据进行基本预处理，发现不准确、不完整、不合理或重复冗余数据并对其进行修补、增减或删除处理，以提高数据质量、保障后续数据分析准确性[12]。数据清洗是整个大数据处理过程中不可或缺的一环，其规范与质量直接关系到随后分析的模型效果和最终结论。在精准医疗领域，数据清洗需要复杂的关系模型，会带来额外的计算成本和延迟开销。如何在大数据清洗模型的复杂性和分析结果的准确性之间做好平衡成为亟待解决的问题。精准医疗领域数据量巨大、增长快速，往往达到 TB 甚至 PB 级存储量，对现有数据清洗工具的工作效率提出较高要求。另外精准医疗涉及数据来源广泛、结构各异，存在不同程度的数据交叉和关联的复杂现象，亟待提高清洗准确率[13]。例如针对多源异构的海量健康医疗数据，需要根据具体数据类型和特点，借助 K－均值聚类、Canopy 算法、K 近邻值、邻近值排序、神经网络、贝叶斯分类等方法，经过数据分析、清洗策略和规则定义、数据校验、数据清洗执行、数据质量评价、干净数据回流等过程，进行缺失、离群、相似或重复、不一致等数据的清洗工作，这不仅对支撑平台的运算能力有较高要求，还需要兼顾数据清洗效率与质量[14－15]。

4.4 数据分析

数据分析用于发现数据中所蕴含的有价值信息，是健康医疗大数据处理流程的核心，也是开展各类精准医疗服务的关键。通过采集、清洗和整合的多源异构数据根据不同精准医疗应用需求，选择部分或全部数据进行集成分析，可实现基于大数据分析的精准医疗服务。精准医疗领域大数据分析需完成庞大的计算量，对处理系统的运算架构、时效性、运算性能和计算域存储单元的数据吞吐率等要求较高，传统分析手段已无法满足大数据环境下的数据分析需求。如何集成现有大数据分析技术，结合精准医疗各类应用的具体需求，研发基于大数据处理综合平台、面向精准医疗服务的大数据分析技术和功能模块，成为亟需解决的问题。以 IBM 的沃森机器人医生为例，为能够向临床医生提供规范化的临床诊疗手段，为患者量身定制个体化治疗和用药方案，提高临床医务人员诊疗质量与服务效率，同时降低医疗事故、不良反应、药物毒副作用等负面事件的发生概率，IBM 为其配备顶尖的计算能力和高效率的自然语言处理技术并构建专业知识库，使其能够每秒处理 500GB 患者临床、实验室检测、病理和生物样本信息等多维度健康医疗数据，从而满足临床辅助决策的实践应用需求[1,16]。

4.5 平台支撑

精准医疗涉及数据繁杂、各类专病应用子模块

众多，且不同专病应用对数据及其处理具有个性化要求。为避免精准医疗服务过程中信息交换规范不统一、专病模块间存在信息孤岛、数据传输不畅等问题，基于平台化技术的数据处理成为未来发展趋势[11,17]。目前在精准医疗应用中，各机构倾向于独自建立数据库和样本库，形成诸多数据烟囱，且大数据处理、隐私保护等对技术与设备条件要求较高，导致进行数据处理的门槛较高[18]。例如高效分布式并行处理大规模多源异构健康医疗数据的平台化模式有 3 种：离线批处理计算框架、流式实时处理计算框架和内存计算框架，部分医疗机构受安全、经济、政治等因素影响而采用一种以上支撑平台技术并集成不同生产厂家大数据处理功能模块，均可对面向精准医疗的大数据处理技术的兼容性、处理效率和质量等造成不同程度影响[11,13]。因此建立面向精准医疗的大数据服务平台，通过云计算、云存储、大数据处理等技术的结合应用，集成数据采集、清洗、融合、质量控制、可视化等功能模块，为精准医疗各类应用提供支撑，成为深入开展精准医疗服务的有效途径。

4.6 数据质量

精准医疗领域海量数据积累迅速，其产生速度远远高于数据分析效率的提升，如何利用大数据处理技术提取有用信息、保证数据质量和分析过程的可重现性成为需要重点考虑的问题。面向精准医疗服务建立规范化、流程化、标准化的大数据质量控制体系，可保证数据质量，提升数据分析效率和数据价值，实现数据对精准医疗服务的有效支撑[2,19]。目前针对精准医疗领域大数据质量控制系统研究较少，缺乏有效的理论框架和技术手段，对精准医疗服务的高效率、高质量发展造成不同程度影响。例如精准医疗领域大数据处理涉及数据采集、清洗、融合、分析、可视化等过程，任何环节的数据处理均可对最终质量造成影响，进而影响数据分析效率与结果准确性，不利于疾病诊断、治疗和用药方案、健康管理措施等精准医疗服务的实施[1-2,19]。

4.7 数据治理

精准医疗领域大数据治理通过协调多个职能部门，基于个性化医疗服务不同目标来制定大数据优化、隐私化、所有权和经营权分配等相关策略，是涉及健康医疗大数据管理、利用、监督和评估的一种支撑保障体系。健康医疗大数据来源广泛、成分复杂、涉及敏感和隐私信息，不能直接使用，必须经过治理方可利用并实现其价值。现有数据治理方法较分散，缺乏整体指导框架，而完善的健康医疗行业数据治理体系尚不成熟，精准医疗领域的数据治理缺乏系统研究。例如患者临床诊疗和日常护理监测过程中产生的健康医疗数据的所有权、使用权、使用规范、隐私保护、利益与责任划分、过程监督等均属于数据治理范畴，目前尚无成熟完善的政策和技术予以支撑。鉴于精准医疗所涉及大数据的来源及特点，结合精准医疗应用现状与需求，为促进精准医疗基于数据驱动的服务创新，有必要构建面向精准医疗的数据治理成套方案。

5 未来研究方向

鉴于健康医疗大数据在精准医疗领域的应用现状与面临问题，未来研究可重点关注以下几个方面。第一，数据采集方面，研发新型大数据采集平台，集成多种数据传输协议和应用程序接口，实现对多源异构健康医疗数据同时统一采集与预处理，根据精准医疗应用实际需求构建专病数据集。第二，数据清洗方面，鉴于健康医疗数据数量巨大、结构多样的特点，可利用深度学习、神经网络等大数据算法对缺失、离群、相似或重复、不一致等数据进行高效率清洗，对数据清洗质量进行评价。第三，数据分析方面，一方面要进行健康医疗大数据标准数据集构建，以提升大数据处理质量与效率；另一方面应积极研发基于健康医疗大数据平台的机器学习算法，以对各种来源的信息进行同时联合分析，从而获得更为可靠、精准、个体化的疾病诊断与治疗辅助决策。第四，平台支撑方面，可研发基于开源 Hadoop 的分布式大数据存储、管理和处理综合服务平台，解决海量健康医疗数据存储、分析与安全管理问题，开发成熟完善的深度学习算法模型，深入挖掘数据蕴含的有价值信息，推动其在疾

病诊断和治疗中发挥积极作用。第五，数据质量方面，可根据精准医疗领域健康医疗大数据处理环节，基于流程视角在数据处理前、数据处理过程中、数据分析后等环节进行质量评估体系构建与实施。第六，数据治理方面，根据精准医疗服务涉及的利益相关方和健康医疗大数据利用过程，进行战略与目标、治理保障、治理域、实施和评估等大数据治理框架功能模块构建与实施。

6 结语

精准医疗是解决我国当前医疗资源紧缺、漏诊误诊率高、医疗费用负担重、药物滥用等医疗卫生领域突出问题的重要途径之一。精准医疗在疾病防治中的应用离不开健康医疗大数据的支撑。本研究通过分析健康医疗大数据在精准医疗领域的应用及挑战，明确健康医疗大数据在精准医疗中的重要作用及应用途径。研究结果对提高健康医疗大数据利用效率，启发未来精准医疗领域大数据相关研究方向，推动精准医疗高速度、高质量发展等具有重要意义。

参考文献

1 詹启敏，张华，陈柯羽，等. 精准医学总论［M］. 上海：上海交通大学出版社，2017.

2 Hulsen T, Jamuar S, Moody A, et al. From Big Data to Precision Medicine［J］. Frontiers in Medicine, 2019（6）: 1 – 14.

3 Song C, Kong Y, Huang L, et al. Big Data – driven Precision Medicine: Starting the Custom – made Era of Iatrology［J］. Biomed Pharmacother, 2020（129）: 110445.

4 Leopold J A, Maron B A, Loscalzo J. The Application of Big Data to Cardiovascular Disease: Paths to Precision Medicine［J］. J Clin Invest, 2020, 130（1）: 29 – 38.

5 Douglass E F. Bridging "Big Data" and Mechanistic Insight to Enable Precision Medicine［J］. Chembiochem, 2020, 21（21）: 3047 – 3050.

6 Cirillo D, Valencia A. Big Data Analytics for Personalized Medicine［J］. Curr Opin Biotechnol, 2019（58）: 161 – 167.

7 Manrai A K, Patel C J, Ioannidis J P A. In the Era of Precision Medicine and Big Data, Who Is Normal?［J］. JAMA, 2018, 319（19）: 1981 – 1982.

8 范美玉，陈敏. 基于大数据的精准医疗服务体系研究［J］. 中国医院管理，2016，36（1）：10 – 11.

9 Primorac D, Bach – Rojecky L, Vadunec D, et al. Pharmacogenomics at the Center of Precision Medicine: Challenges and Perspective in an Era of Big Data［J］. Pharmacogenomics, 2020, 21（2）: 141 – 156.

10 Pastorino R, De Vito C, Migliara G, et al. Benefits and Challenges of Big Data in Healthcare: an Overview of the European Initiatives［J］. Eur J Public Health, 2019, 29（Suppl 3）: 23 – 27.

11 高景宏，翟运开，何贤英，等. 面向精准医疗的大数据采集及其支撑要素研究［J］. 中国卫生事业管理，2020，37（6）：405 – 407，425.

12 杨尚林. 基于机器学习的多源异构大数据清洗技术研究［D］. 南宁：广西大学，2017.

13 高景宏，赵杰，李明原，等. 面向精准医疗的多源异构数据融合技术研究［J］. 医学信息学杂志，2021，42（5）：69 – 74.

14 韩帅，孙乐平，杨艺云，等. 基于改进 K – Means 聚类和误差反馈的数据清洗方法［J］. 电网与清洁能源，2020，36（7）：9 – 15.

15 毛云鹏，龙虎，邓韧，等. 数据清洗在医疗大数据分析中的应用［J］. 中国数字医学，2017，12（6）：49 – 52.

16 Zhang X, Perez – Stable E J, Bourne P E, et al. Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century［J］. Ethn Dis, 2017, 27（2）: 95 – 106.

17 翟运开，武戈. 基于电子病历信息大数据挖掘的患者就医行为分析［J］. 医学信息学杂志，2017，38（7）：12 – 17.

18 Prosperi M, Min J S, Bian J, et al. Big Data Hurdles in Precision Medicine and Precision Public Health［J］. BMC Med Inform Decis Mak, 2018, 18（1）: 139.

19 Hopp W J, Li J, Wang G. Big Data and the Precision Medicine Revolution［J］. Production and Operations Management, 2018, 27（9）: 1647 – 1664.