·学者论坛·

生物医学大数据:现状与展望

王波 吕筠 李立明

【关键词】 生物医学; 大数据; 个体化医学

Big data in biomedicine: status quo and perspective *Wang Bo*[1], *Lyu Jun*[2], *Li Liming*[2]. 1 *Health Science Popularization Research Center, Chinese Academy of Medical Sciences, Beijing* 100730, China; 2 *School of Public Health, Peking University*
Corresponding author: Li Liming, Email: lmlee@vip.163.com

【Key words】 Biomedicine; Big data; Personalized medicine

大数据(big data)是指由于容量太大和过于复杂,无法在一定时间内用常规软件对其内容进行抓取、管理、存储、检索、共享、传输和分析的数据集。大数据具有"4V"特征:①数据容量(Volume)大,常常在PB(1 PB=250 B)级以上;②数据种类(Variety)多,常常具有不同的数据类型(结构化、半结构化和非结构化)和数据来源;③产生和更新速度(Velocity)快(如实时数据流),时效性要求高;④科学价值(Value)大,尽管利用密度低,却常常蕴藏着新知识或具有重要预测价值[1]。人类已进入大数据时代。国际数据公司的研究结果表明,2011年全球产生的数据量高达1.82 ZB[2]。2012年5月,联合国发布了《大数据与人类发展:挑战与机遇》白皮书,指出大数据是一个历史性机遇,人们可以使用极为丰富的数据资源对社会经济进行前所未有的实时分析,帮助政府更好地响应社会和经济运行[3]。

大数据受到越来越多的重视。欧美国家许多高校纷纷成立了数据科学研究机构,开设了数据科学课程。*Nature*和*Science*也分别于2008年和2011年推出了大数据专刊,对大数据带来的挑战进行讨论[4]。作为最活跃的科学研究领域之一,生物医学领域的大数据也备受关注。

1. 生物医学大数据的来源:以下因素促进了生物医学领域大数据的出现。①生命的整体性和疾病的复杂性。例如,严重威胁人类健康的慢性病多为复杂性疾病,其发生具有复杂的遗传和分子机制,受到基因、环境及其交互作用的影响,其病因学研究将产生大量的数据[5]。②高通量技术的发展和基因组测序成本的下降。高通量测序技术可以对数百万个DNA进行同时测序,使得对一个物种的转录组和基因组进行细致全面的分析成为可能[6]。随着人类基因组计划的完成和计算能力的快速发展,每个基因组的测序成本已从数百万美元降低至数千美元(并且还将继续降低)[7]。这将产生海量测序数据(每个人的基因组就需要3 G的数据存储量)。③医院信息化和IT业的迅速发展。人体本身就是生物医学大数据的一个重要来源,随着医院信息化和IT业的迅速发展,越来越多的人体数据能够获得储存和利用。例如,X线、3D核磁、乳腺X线、3D CT扫描分别包括30 M、150 M、120 M和1 G的数据量,至2015年美国平均每家医院需要管理665 T的数据量[8]。

生物医学大数据广泛涉及人类健康相关的各个领域:临床医疗、公共卫生、医药研发、医疗市场与费用、个体行为与情绪、人类遗传学与组学、社会人口学、环境、健康网络与媒体数据(表1)[9]。

表1 生物医学大数据的主要来源

数据来源	具体类型
临床医疗	电子病历、医学影像、医疗设备监测等
公共卫生	疾病与死亡登记、公共卫生监测、电子健康档案、食品销售、营养标签等
医药研发	临床试验、药物研发、医疗设备研发等
医疗市场与费用	医疗服务费用、医疗设备销售记录、药店销售记录、医疗保险等
个体行为与情绪	实时视频、个体行为、健身记录、体力活动记录、缺勤记录、传感器等
人类遗传学与组学	基因组学、转录组学、蛋白质组学、代谢组学等
社会人口学	性别、年龄、婚姻状况、经济收入等
环境	休闲场所、污染、犯罪、交通等
健康网络与媒体	健康网站、搜索引擎、通讯运营商、微博、微信、论坛、即时通讯等

2. 生物医学大数据的应用:生物医学大数据可应用于以下方面。①开展组学研究及不同组学间的关联研究。从环境、个体生活方式行为等暴露组学,

DOI:10.3760/cma.j.issn.0254-6450.2014.06.001
作者单位:100730 北京,中国医学科学院健康科普研究中心(王波);北京大学公共卫生学院(吕筠、李立明)
通信作者:李立明,Email:lmlee@vip.163.com

至个体细胞分子水平上的基因组学、表观组学、转录组学、蛋白组学、代谢组学、宏基因组学，再到个体健康和疾病状态的表型组学等。利用大数据将各种组学进行综合及整合，既能为疾病发生、预防和治疗提供全面、全新的认识，也有利于开展个体化医学，即通过整合系统生物学与临床数据，可以更准确地预测个体患病风险和预后，有针对性地实施预防和治疗[7,10]。②快速识别生物标志物和研发药物。利用某种疾病患者人群的组学数据，可以快速识别有关疾病发生、预后或治疗效果的生物标志物[11]。在药物研发方面，大数据使得人们对病因和疾病发生机制的理解更加深入，从而有助于识别生物靶点和研发药物；同时，充分利用海量组学数据、已有药物的研究数据和高通量药物筛选，能加速药物筛选过程[7]。③快速筛检未知病原和发现可疑致病微生物。通过采集未知病原样本，对病原进行测序，并将未知病原与已知病原的基因序列进行比对，从而判断其为已知病原或与其最接近的病原类型，据此推测其来源和传播路线、开展药物筛选和相应的疾病防治[12]。④实时开展生物监测与公共卫生监测。公共卫生监测包括传染病监测、慢性非传染性疾病及相关危险因素监测、健康相关监测（如出生缺陷监测、食品安全风险监测等）。此外，还可以通过覆盖全国的患者电子病历数据库进行疫情监测[13]，通过监测社交媒体或频繁检索的词条来预测某些传染病的流行[14]。例如，Google Trends通过找寻“流感症状”和“流感治疗”之类搜索词的峰值，在医院急诊流感患者增加之前就能对某些地区的流感做出预测[15,16]。⑤了解人群疾病谱的改变。这有助于制定新的疾病防治策略。全球疾病负担研究是一个应用大数据的实例[17]，该研究应用的数据范围广、数据量巨大，近4 700台并行台式计算机完成了数据准备、数据仓库建立和数据挖掘分析的自动化和规范化计算。其有关中国的研究发现：与1990年相比，2010年造成中国人群寿命损失的前25位病因中，慢性非传染性疾病显著上升，传染病则显著下降，说明慢性非传染性疾病已经成为我国人群健康的主要威胁[18]。⑥实时开展健康管理。通过可穿戴设备对个体体征数据（心率、脉率、呼吸频率、体温、热消耗量、血压、血糖、血氧、体脂含量等）的实时、连续监测，提供实时健康指导与建议，更好地实施健康管理[9,19,20]。⑦实施更强大的数据挖掘。数据挖掘的任务包括关联分析、聚类分析、分类分析、异常分析等。大数据挖掘能够增加把握度和发现弱关联的能力[13]。

3. 生物医学相关的大数据计划：近年来国内外一些生物医学相关的大数据计划见表2。

4. 生物医学大数据面临的主要问题与发展趋势：作为一个新兴领域，大数据也伴随着一些争议[27,28]：①既然数据总是不断增加，是否有必要区分大数据与传统数据？②大数据更多意义上可能是一种商业上的宣传？③大数据中变量类型更多、更复杂，而随着变量的增加，获得假阳性关联的概率也会增加；④更大的数据未必意味着更好的数据，必须考虑数据的代表性和数据纯度；⑤在未告知个体的情况下使用来自人群的数据是否符合伦理学要求？这些争议是大数据在未来发展中必须关注的。

从流行病学角度来看，生物医学大数据具有以下优势：①具有大样本的特点，能够解决流行病学研究中的样本量问题，大样本能够提高结果精度高、降低随机/抽样误差；②客观的采集途径能够减少信息偏倚。大数据的采集途径往往比较客观，还能全程动态地记录个体行为，相比传统流行病学调查通过询问、回忆某些行为的状况，能够减少信息偏倚。然而，相对于传统概率随机抽样而言，大数据

表2　近年生物医学相关的大数据计划

名　称	制定机构（发布时间）	主要目的	生物医学相关的内容	产出代表
大数据研究和发展计划[21,22]	美国政府（2012年3月）	提高从大数据中获取知识和观点的能力，用以解决国家面临的重大挑战	发布生物医学大数据科研招标资助以确定蛋白结构和生物学路径为目的的研究	国际千人基因组计划
“从数据到知识再到行动”[23]	美国政府（2013年11月）	可视为大数据研究和发展计划的第二期	疾病大流行预测项目“从大数据到知识”计划	癌症基因组图谱、心血管研究网络
“全球脉动”[24]	联合国（2009年）	通过早期预警、实时反馈，更及时地追踪和监测全球和地区社会经济危机	预测疾病暴发 监测食品安全问题	全球脉动实验室网络
人口与健康科学数据共享平台[25]	中国科技部（2010年）	通过科学数据汇交、数据加工、数据存储、数据挖掘和数据共享服务，为创新型人才培养和健康产业发展提供科学数据共享服务	整合全国健康领域数据资源提供健康数据共享服务	脑卒中专题服务、农村卫生专题服务
上海推进大数据研究与发展三年行动计划[26]	中国上海市科技委员会（2013年7月）	加速大数据资源的开发利用，支撑智慧城市建设	建立全民医疗健康服务平台 建立食品安全大数据服务平台	–

可能存在选择偏倚问题,其收集途径常常覆盖的是具有某些特征的人群(如医保患者、使用可穿戴设备的人群)。

生物医学大数据面临的主要问题:①如何实现生物医学数据的标准化和规范化。数据标准化是数据共享的前提,只有标准化的数据才能有效融合与整合,从而发挥大数据的价值[29]。②如何打破数据孤岛,实现生物医学数据共享[25]。应避免数据只为少数人或单位使用,数据共享是应用生物医学大数据的前提。许多公共资助机构已开始要求所资助研究的数据必须在一定范围内共享[30]。③生物医学大数据的存储和管理。生物医学领域数据特别庞大,产生和更新速度更快,其存储方式不仅影响数据分析效率,也影响数据存储的成本[4]。④如何实现生物医学大数据的高效利用。我国已积累了海量的生物医学数据,如何利用才是关键,这在一定程度上也依赖于大数据技术的发展[31]。⑤生物医学大数据的分析、整合与挖掘。特别是对半结构化和非结构化数据(如心电图、医学影像资料)和对流数据(实时视频、传感器数据、医疗设备监测数据)的处理,是生物医学大数据分析面临的重要挑战[32]。⑥生物医学和信息科学的复合型人才缺乏。这是国内外生物医学大数据面临的一个困境,需要推动计算机科学和生物学交叉学科的教育予以解决[4,17]。

未来生物医学大数据的发展趋势[33]:①从"概念"走向"价值",成为"智慧健康"的基础。生物医学大数据将能够产生新的知识,用信息改变医学实践,最终改善人类健康和公共卫生[11]。②医学科学证据的整合、转化和循证科学证据的产生。生物医学大数据有助于循证科学证据的生产,例如通过大数据可以对大量健康数据进行整合,进而获得更加可靠的证据[9];还可以通过网络实时数据,开展"虚拟的临床试验"生产证据[34]。③数据安全与隐私保护的技术发展。在对海量数据进行挖掘的同时,隐私泄露存在巨大风险。数据安全与隐私保护日益受到关注和重视,相关政策和立法亟待加强,相应的技术发展将发挥重要作用[9]。④大数据为导向的人群队列研究成为热点。超大规模队列研究具有大样本(数十万人群)、前瞻性(数十年长期随访)、多学科(基础、临床、预防、信息等多学科合作)、多病种(能够对多种疾病进行研究)、多因素(能够探讨多种危险因素)、整合性(监测系统、信息系统、医保系统的整合)、共享性(生物标本和数据资源的共享)等特点,经过长期随访能够产出大量人群数据[35]。⑤生物医学大数据的可视化。可视化与信息图像、信息可视化、科学可视化以及统计图形密切相关,能够更清晰有效地传达与沟通大数据包含的信息[13,36]。⑥基于生物医学大数据的个体化健康管理逐步流行。一方面,利用实时的传感器(可穿戴设备),能够对个体进行实时的、连续的健康监测与评估,为个体提供实时健康指导[31];另一方面,随着以生物医学大数据为基础的个体化医学发展,个体化预防、诊断和治疗将得以实现[11]。⑦生物医学大数据成为战略性产业。许多国家已经将大数据上升为国家层面战略,生物医学大数据产业化已经初现[4]。

5. 展望:人类已进入大数据时代。大数据科学作为一个横跨信息科学、社会科学、网络科学、系统科学、生物医学、心理学、经济学等诸多领域的新兴交叉学科方向正在逐渐形成,并已成为科学研究热点[4]。生物医学领域具有海量数据,如何共享、规范、管理和利用是关键。同时,生物医学大数据专业人才培养亟待解决。生物医学大数据将改变医学实践模式,改善医药卫生服务质量,最终有利于实现个体化治疗和群体性预防的医学目的。

参 考 文 献

[1] Jee K, Kim GH. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system [J]. Healthc Inform Res, 2013, 19(2): 79-85.

[2] Mayer-Schönberger V, Cukier K. Big Data: a revolution that will transform how we live, work, and think [M]. Boston: Houghton Mifflin Harcourt, 2013.

[3] UN Global Pulse. Big data for development: challenges and opportunities [EB/OL]. (2012-05-01) [2014-04-09]. http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-GlobalPulseMay2012.pdf.

[4] Li GJ, Cheng XQ. Research status and scientific thinking of big data [J]. Bull Chin Academy Sci, 2012, 27(6): 647-657. (in Chinese)
李国杰,程学旗. 大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊,2012,27(6):647-657.

[5] Hu YH. Population based etiological research on complex diseases [J]. J Peking University: Health Sci, 2007, 39(2): 113-115. (in Chinese)
胡永华. 基于群体的复杂性疾病的病因研究[J]. 北京大学学报:医学版,2007,39(2):113-115.

[6] Wang XC, Yang ZR, Wang M, et al. High-throughput sequencing technology and its application[J]. Chin Biotechnol, 2012, 32(1): 109-114. (in Chinese)
王兴春,杨致荣,王敏,等. 高通量测序技术及其应用[J]. 中国生物工程杂志,2012,32(1):109-114.

[7] Costa FF. Big data in biomedicine[J]. Drug Discov Today, 2014, 19(4): 433-440.

[8] Monty Zarrouk, NetApp. Delivering excellence in patient care with ready access to clinical data [EB/OL]. (2012-09-01) [2014-04-09]. http://www.netapp.com/us/media/wp-7169.pdf.

[9] Groves P, Kayyali B, Knott D, et al. The Big Data Revolution in Healthcare: Accelerating Value and Innovation [M]. New York (NY): McKinsey Global Institute, 2013.

[10] Murdoch TB, Detsky AS. The inevitable application of big data to health care[J]. JAMA, 2013, 309(13): 1351-1352.

[11] Chawla NV, Davis DA. Bring big data to personalized healthcare: a patient-centered framework [J]. J Gen Intern Med, 2013, 28 Suppl 3: S660-665.

[12] Yang RF. Preventive medicine research in the era of big data: digital preventive medicine[J]. Chin J Prev Med, 2014, 48(3): 1-4. (in Chinese)
杨瑞馥. 大数据时代的预防医学研究:数字化预防医学[J]. 中华预防医学杂志, 2014, 48(3): 1-4.

[13] Gao HS, Xiao L, Xu DW, et al. Medical data mining platform based on cloud computing[J]. J Med Inform, 2013, 34(5): 7-12. (in Chinese)
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.

[14] Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak[J]. Am J Trop Med Hyg, 2012, 86(1): 39-45.

[15] Dugas AF, Jalalpour M, Gel Y, et al. Influenza forecasting with Google Flu Trends[J]. PLoS One, 2013, 8(2): e56176.

[16] Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks[J]. Clin Infect Dis, 2009, 49(10): 1557-1564.

[17] Yu SC, Xiao GX. Research on global burden of disease-big data analysis application examples[J]. J Med Inform, 2013, 34(9): 12-16. (in Chinese)
于石成, 肖革新. 全球疾病负担研究——大数据分析应用实例[J]. 医学信息学杂志, 2013, 34(9): 12-16.

[18] Yang G, Wang Y, Zeng Y, et al. Rapid health transition in China, 1990-2010: findings from the Global Burden of Disease Study 2010[J]. Lancet, 2013, 381(9882): 1987-2015.

[19] Xu DQ, Yang HQ. The application of big data on healthcare personalized service[J]. Chin J Health Inform Manag, 2013, 10(4): 301-304. (in Chinese)
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.

[20] Zhou GH, Xin Y, Zhang YJ, et al. Study on big data's application in medical and health field[J]. Chin J Health Inform Manag, 2013, 10(4): 296-300, 304. (in Chinese)
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, 304.

[21] Office of Science and Technology Policy, Executive Office of the President of the United States. Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments [EB/OL]. (2012-03-29) [2014-04-09]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

[22] Siva N. 1000 Genomes project[J]. Nat Biotechnol, 2008, 26(3): 256.

[23] Executive Office of the President of the United States. Fact Sheet: Data to Knowledge to Action [EB/OL]. (2013-11-12) [2014-04-09]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/Data2Action%20Press%20Release.pdf.

[24] UN Global Pulse [EB/OL]. [2014-04-09]. http://www.unglobalpulse.org.

[25] Jiang JD, Zhao H, Liu RD. Discussion on the information organization of the scientific data sharing platform-take the National Scientific Data Sharing Platform for Population and Health as an example[J]. J Inform Resources Manag, 2012(4): 52-56. (in Chinese)
姜吉栋, 赵辉, 刘润达. 科学数据共享平台网站中的信息组织:以国家人口与健康科学数据共享平台为例[J]. 信息资源管理学报, 2012(4): 52-56.

[26] Shanghai Municipal Science and Technology Commission. Big data research and development in Shanghai: a three-year action plan (2013-2015) [EB/OL]. (2013-07-12) [2014-04-09]. http://www.stcsm.gov.cn/gk/ghjh/333008.htm. (in Chinese)
上海市科学技术委员会. 上海推进大数据研究与发展三年行动计划(2013-2015年)[EB/OL]. (2013-07-12)[2014-04-09]. http://www.stcsm.gov.cn/gk/ghjh/333008.htm.

[27] Chiolero A. Big data in epidemiology: too big to fail?[J]. Epidemiology, 2013, 24(6): 938-939.

[28] Fan W, Bifet A. Mining big data: current status, and forecast to the future[J]. SIGKDD Explorations, 2012, 14(2): 1-5.

[29] Lynch C. Big data: how do your data grow?[J]. Nature, 2008, 455(7209): 28-29.

[30] National Institutes of Health Office of Extramural Research. NIH Sharing Policies and Related Guidance on NIH-Funded Research Resources [EB/OL]. [2014-04-09]. https://grants.nih.gov/grants/sharing.htm.

[31] Trifonova OP, Il'in VA, Kolker EV, et al. Big Data in Biology and Medicine[J]. Acta Naturae, 2013, 5(3): 13-16.

[32] Green DE, Rapp EF. Can big data lead us to big savings?[J]. RadioGraphics, 2013, 33(3): 859-860.

[33] China Computer Federation. Trends of big data in 2014 [J]. Communic CCF, 2014, 10(1): 32-36. (in Chinese)
中国计算机学会大数据专家委员会. 2014年大数据发展趋势预测[J]. 中国计算机学会通讯, 2014, 10(1): 32-36.

[34] Magid DJ, Gurwitz JH, Rumsfeld JS, et al. Creating a research data network for cardiovascular disease: the CVRN[J]. Expert Rev Cardiovasc Ther, 2008, 6(8): 1043-1045.

[35] Xiong WY, Lv J, Guo Y, et al. Overview on the practice and characteristics of large prospective cohort studies[J]. Chin J Epidemiol, 2014, 35(1): 93-96. (in Chinese)
熊玮仪, 吕筠, 郭彧, 等. 大型前瞻性队列研究实施现况及其特点[J]. 中华流行病学杂志, 2014, 35(1): 93-96.

[36] Callebaut W. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology [J]. Stud Hist Philos Biol Biomed Sci, 2012, 43(1): 69-80.

(收稿日期:2014-04-09)
(本文编辑:王岚)