·综述·

健康大数据的来源与应用

曲翌敏 江宇

【关键词】 大数据；健康；来源；应用

The sources and application of big data in healthcare *Qu Yimin, Jiang Yu. Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing* 100730, China

Corresponding author: Jiang Yu, Email: wingedsky@gmail.com

This work was supported by grants from the PUMC Youth Fund and the Fundamental Research Funds for the Central Universities (No. 3332015015).

【Key words】 Big data; Healthcare; Source; Application

1. 健康大数据的定义与来源：大数据是社会信息化发展的必然产物，对其具体定义，众说纷纭，尚无统一答案。《大数据时代》一书中，大数据是指不采用抽样这一调查方法，而是对所有数据进行分析处理[1]。Gartner公司对于大数据的定义：又称巨量资料，是指需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。而笔者更认同互联网周刊所述观点，即“大数据”远不止大量的数据和相应的处理技术，或所谓的“4V”简单概念，而是包括了人们在大数据的基础上可以做的事情，这些事情在小规模数据的基础上无法实现。大数据让我们以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见，最终形成变革之力[2]。

大数据与健康领域的结合目的正是如此，即通过对健康相关大数据的利用，深入、有效地进行流行病的病因学探讨，辅助临床治疗与决策，帮助卫生政策的制定，服务于人类的健康改善。可穿戴设备、新一代基因测序等新科技的发展与卫生统计数据和病案数据等的电子化，使得健康相关数据呈指数型增长，数据的可及性、可利用性提高，数据价值得以实现[1,3]。健康大数据来源广泛，形式复杂多样，主要包括6类。

（1）个人日常生理数据：“可穿戴设备”迅速发展，产品形态多样，功能逐渐丰富，健康相关手机应用更是迅速增多[4]。随着其受众范围的不断扩大，人体健康数据的采集变得轻而易举，所有佩戴者都成为了传感器，不断地收集与传递自身的健康数据。

（2）基因组学、蛋白质组学数据：第三代与第四代DNA测序技术的发展，使迅速而价格相对低廉的基因测序成为可能，千人基因组计划与DNA元件百科全书计划（ENCODE）是其中两个重要代表。前者在已完成的第一阶段就已经对超过1 000个基因组进行了测序，人类一个基因组测序结果大约占140 GB[5]，该项目产生的数据库是目前关于人类的遗传变异的最大数据库，其成果被广泛应用。后者在大约150种细胞类型中进行了1 600次试验，这些试验产生了惊人的数据与信息[6]。在中国，深圳华大基因研究院每天约产生6 TB的基因组学数据[7]。2015年1月30日，美国总统奥巴马宣布了一项名为“精准医学”的计划，这项计划的核心在于创建一个囊括各个年龄阶层、各种身体状况志愿者的基因数据库，这些数据将为癌症及多种疾病的个性化治疗提供有价值的信息。

（3）医院电子病例数据、电子健康档案数据：现代医疗系统每天的运作会产生大量的电子数据，包括电子病历数据、医学检验数据与医学影像数据，据统计，一家三甲医院大约每天会产生几十TB的数据[8]，而且随着新技术的不断引入，电子病历系统的不断完善，这个数字还会继续增大。20世纪90年代中后期，随着对电子病历系统化研究的日益深入，西方发达国家纷纷开始致力于对电子健康档案的研究[9]。我国是从2009年开始在全国范围建立统一的居民健康档案，并实施规范管理，健康档案更符合以健康为中心的卫生服务模式，能够提供覆盖全人群准确的、全面的个人健康信息及就诊记录数据，有助于医生的诊疗活动与社会的慢病防控与健康教育工作。

（4）大型队列研究、医疗科研数据：随着20世纪中期以来对慢性病研究的不断深入，大型队列研究逐渐兴起，其中最著名的Framingham的心血管病队列研究历经三代人60余年，其研究成果对于心血管病的防治具有里程碑意义，至今指导着心血管病防治工作[10]。现代队列研究已发展到数十万甚至数百万人的规模，如中国慢性病前瞻性研究项目（China Kadoorie Biobank, CKB），是由中国医学科学院与牛津大学合作的一项涉及中国10个地区51万余人、将持续长达20～30年的前瞻性队列研究，旨在研究危害中国人群健康重大慢性病的致病因素、发病机理、流行规律和趋势，为有效地制定慢性病预防和控制对策，开发新的治疗和干预手段提供科学依据[11]。这些大型队列研究最直接的产出便是不容小觑数据量，而研究意义与价值全部隐藏在这些数据当中等待被进一步挖掘。临床试验也会产生大量的数据，新药研发从基础研究到上市平均耗时15年，耗费大量人力物力，最直接的产出便是试验数据，这些数据必须实现共享与长远利

DOI: 10.3760/cma.j.issn.0254-6450.2015.10.031

基金项目：协和青年基金；中央高校基本科研业务费专项资金（3332015015）

作者单位：100730 北京，中国医学科学院/北京协和医学院公共卫生学院

通信作者：江宇，Email：wingedsky@gmail.com

用才会达到最大效益，对公众健康产生最大化利益[12]。1997年，在《FDA现代化法》的规定下，Clinicaltrials.gov网络搭建起来，至2015年4月29日，来自190个国家的189 109项临床试验进行了登记[13]。

（5）网络健康数据：各种健康网站、手机app不断涌现，会员人数急剧增加，不同的网站或手机app面向不同的病患群体，提供不同的服务，如用药指导、健康管理、提供快捷方便的就诊环境、或搭建病友相互交流的平台等，搜集到了大量的患者健康数据，可以用来为疾病或药品的研究提供服务。除此之外，普通社交网络中也隐藏着大量的疾病与健康分享数据[14-15]。各大搜索引擎则收集了人们的大量“行为数据”，人们习惯在网上搜索疾病的治疗方法与危险因素，或者保健养生之道等。总之，健康数据零散分布在互联网的各个角落等待挖掘与利用。

（6）公共卫生数据：公共卫生数据主要是指疾病监测与卫生监督数据，包括传染病、慢性病、症候群及病原的监测以及餐饮、食品、水源的监测。其包含范围较广，由多渠道所得多种数据相融合形成一个庞大的信息系统，如美国疾病预防控制中心（CDC）生物监测系统与我国的中国疾病预防控制信息系统。

2. 健康大数据的应用领域：大数据的兴起改变了人们以往对数据的看法，在上述健康大数据的来源中我们可以看到，除了对传统的交易数据（如病例数据、研究数据）的分析利用之外，行为数据（日常监测数据、网站浏览数据）也逐渐引起了研究者们的重视。数据中隐藏着高价值的信息，2012年美国政府开始拉动大数据相关产业发展，并将“大数据战略”上升至国家战略。麦肯锡在《大数据：下一个竞争、创新和生产力的前沿领域》研究报告中指出，大数据在医疗领域每年能够产生3 000亿美元的潜在价值[16]。通过对健康数据的挖掘，将得到的信息用于医学研究、流行病预测、药物副作用分析等方面，可以帮助推动医疗技术的发展，提高居民健康水平。对于大数据在医疗领域的应用期待现主要聚焦在如下领域中，当然，对于能在其他领域中带来的颠覆，我们拭目以待。

（1）聚焦公共卫生大数据应用：坚实的流行病学基础，强大的知识整合能力，以及循证医学的原则，将促进大数据在公共卫生领域的应用[17]。首先是公共卫生预警与疾病预测：大数据应用核心在于预测，在2014年的Ebola的疫情控制中，CDC的专家利用流行病学数据建立马尔可夫链模型预测发病率，强有力地向高层决策者揭示了后果的严重性，促进了资源调动与迅速的应急行动[18-19]。谷歌通过对美国几十亿条互联网检索数据分析，找到了一个能够推测出某个城市的流感状况数学模型。有研究者通过社交网络数据的挖掘对HIV患者进行远程监测，其结果与CDC的监测结果高度一致，且基于网络行为数据的预测更经济、及时[20]。大数据在公共卫生预警方面给我们带来了前所未有的机遇[1]。在过去十年里大数据也已成功应用于心脏病的预测、肝癌特征的识别当中，相信未来大数据可以对于慢性病的预测提供更有价值的帮助[21]。其次，健康大数据给流行病学寻找病因线索提供了新的途径：对电子病历与健康档案中的数据挖掘可以不仅对医疗产品、干预进行评价，改善医疗行为，支持临床决策[22]，而且每次就诊都会给该系统注入新的信息，这可以创造出“学习型医疗系统”，帮助流行病学专家进行病因探讨，揭示未知的疾病与暴露相关关系[23-24]。此外，对于罕见的暴露或结局的研究，通过对患者常规健康资料的数据挖掘和监测（基于大型数据库的队列研究）可将不同来源的数据进行融合统计分析，解决随机对照试验与队列研究等前瞻性设计在探讨慢性罕见治疗不良反应上的大样本需求难以满足的缺陷，且不同来源的数据往往基本人口特征不同，使我们可以进行更全面的分析，实时监测数据则更有利于危险因素早发现与积极防控。全基因组测序和大数据的融合，使我们对传染病传播与控制的认识模式发生转变，但如果过度解读这些数据，也可能会存在潜在的陷阱，使我们找到似是而非的感染途径[25]。

（2）提高诊疗质量：人与人之间的差异是巨大的，年龄、性别、体重、新陈代谢速率和基因变异的类型，综合考虑这些因素后会发现，治疗的标准计量并不适用于每位患者。药效在人群中反应不一，这一事实在20世纪90年代末就已经被发现，但由于缺少患者的整体健康数据，医生没有办法根据每位患者的具体情况来制定预防措施或治疗方案。但在大数据时代，我们可以通过基因组学、无线传感器、数字化成像和医疗信息技术的联合应用，对每位患者的健康数据进行整合和分析，得到该患者家族遗传史、基因组情况、用药史等数据，发现患者疾病的早期标记物，测量患者的患病危险度，提示目前可能存在的隐患，提供早期诊断与预防参考与保健策略，对该患者的就诊进行个体化指导[26]。另一方面，通过检索数据库中拥有共同特征患者的疾病机制、病因与治疗方案与预后信息，可以帮助大夫更好地把握疾病的诊疗。我们相信，将大数据融入到医疗诊断中建立临床决策支持系统，可以提高诊疗与护理的质量，降低二次住院率，促进个体与群体健康状况的改善，推动个体化医疗的进一步发展[27]。由生物标记、治疗方案和无线传感器三者组成的统一体，将成为推动未来医学发展的强大动力。此外，利用移动医疗进行健康管理，可以帮助健康人群提高健康素养，改善健康行为；帮助危重患者或者年迈患者的健康与医院建立无间断、实时的连接，从而得到很好的监护；帮助慢性病患者控制危险因素、提高依从性。对于改善疾病管理模式，帮助医生实现疾病的社区管理，提高人民生活水平有重要的意义[28-29]。

（3）促进医学科学研究与药物研究：在传统的临床科研架构下，医生的临床工作与临床研究往往是分离的，临床工作者想要进行一项研究时往往需要从头开始收集数据，而科研项目又存在资金与时间限制，科研项目往往不能长期地做下去，但疾病尤其慢性病的转归需要较长时间，使得临床工作者的临床科研很难有一个长期地规划与实施，阻碍着临床科研的发展。健康大数据的崛起为这种状况提供了改善机会，通过将临床工作中逐步、长期收集起来的大量数据在研

究中运用[30]，实现临床工作和科研一体化。科研迅速发展的今天，每个人的科研结果都要以别人的科研结果为铺垫或支撑，研究者们越来越依赖于他人的研究数据。科研数据的共享与利用影响着科学研究的快速发展[31]。尤其循证医学发展到今天，临床医生必须实时了解最新的临床实践指导证据，医学文献的数字化使之成为可能[32]。

（4）大数据提供决策支持：意大利大学联盟（CINECA）建立了一个以人群为基础、以患者为中心的数据库，并利用该数据库监测该国的卫生经济情况、患者健康情况以及衡量卫生服务执行情况。该数据库自1987年开始定期对1 200万人口的国家健康系统中的数据与其他研究的高质量完备的数据整合，提供患者的基本人口学特征、院外用药情况、住院费用清单、影像与实验室检查等数据。利用这些数据我们可以对整体或其中几个队列进行慢性病患病率、总发病率与疾病经济负担进行分析。一个完备地将卫生行政数据与临床数据进行整合的大数据平台，是卫生管理与决策支持的有效工具[33]。

3. 健康大数据发展中存在的问题：大数据的本质在于对数据的核心价值进行挖掘，其Volume（数据量大）、Velocity（数据更新快、时效性高）、Variety（数据类型多）、Value（数据价值密度低、应用价值高）的“4V”特征[1]决定了对大数据的有效利用更非易事。相比其他学科的大数据，健康大数据的来源更加广泛，结构更为复杂[24]。对于健康大数据的利用现在还存在着很多问题，有着很大的发展空间。

（1）结构化数据向非结构化数据的转变：数据正在变得无处不在、触手可及，而数据能够创造的真正价值取决于数据整合与分析[21]。将数据处理为信息，再提取为可以应用的知识，最终用于解决实际问题形成智慧，数据量不断减少，获取成本逐渐升高，中间存在着“知识鸿沟”与“智慧鸿沟”，如何通过数据的整理与分析，跨越这两大鸿沟是大数据利用的核心问题[34]。传统的数据分析是基于清晰合理的数据，数据分析模式通常为假设——验证假设，致力于找寻因果关系。但是大数据时代对数据的精确性要求降低[35]，非结构化数据占总数据的70%～80%，传统的统计分析方法无法对其进行深入的分析[36]，数据分析追求的是效率和相关关系。更大的样本量有时并不等同于更可靠的结果，部分数据分析导向的流行病学研究颠倒了提出问题——数据分析的研究思路[37]，先通过数据分析技术寻找相关关系，再来解释这种关联，使得结果为虚假关联的可能性很大[38-39]。健康领域不同于商务领域，不管是寻求疾病的危险因子，还是某种诊疗的效果评估，模糊的相关关系都是不够的[17]，我们都希望得到准确的、能够解释的因果关系，才能采取进一步的防治或改善措施[27]。其次，健康大数据的分析应该是根据数据的更新自动进行，对于及时性的要求更高[7]。对于“可穿戴设备”收集到的实时持续生理数据，需要高效、自动、准确地分析处理和回应。

（2）信息孤岛的问题：现今数以万计数的公立、民营医院各自独立设计运行着互不兼容的电子化医疗档案系统，不同系统之间的信息不能实现完全互通，再加上隐私、安全、监管与专利等诸多因素限制着数据共享[30]，使得电子病历系统成了一个个的“信息孤岛”，患者连续的健康信息无法被综合利用。如果能将各医院、区域、城市，甚至国家间的医疗数据系统进行统一的规范后搭建区域医疗集中平台，建立医疗健康服务体系，将有助于提高诊疗质量、控制医疗费用以及方便患者的转诊[40-41]。想要打破“信息孤岛”局面，需要我们对数据的格式建立统一行业标准，对数据源进行统一定义，以大医院带动小医院搭建一个过程可互通的健康数据平台[40,42]。

（3）数据存储安全性、保密性与管理：数据要能被有效利用，首先要确保有足够的计算机设备来存储、运行这些健康大数据。随着数据容量的指数型增长，对于数据存储空间不足这一问题主要靠购买商业化云存储空间来解决。很多医学试验大数据的产生往往是世界各地的各大实验室共同工作的结果，对于这类试验来说，数据的传输与分享尤为重要，云存储和云计算显示出了巨大的优势，但也引入了新的安全性问题：数据的存储安全、个人健康信息的保密。数据的有效管理是大数据可及、可信、安全的重要保障[42-43]。

4. 建议：健康大数据是加速医疗与公共卫生发展的一个重要工具，如何让健康大数据的价值充分实现，需要我们做的工作还很多，顶层的宏观设计、把握，政策环境的支持以及良好的激励机制，是促使健康大数据相关的各行业部门打破壁垒，解决“信息孤岛”问题的保障。跨领域人才队伍紧密配合、专业的知识与技术是数据有效分析与数据安全、个人隐私得以保障的实现基础[19]。

最后我们还应注意，任何一门技术都有其优缺点，大数据固然给健康领域带来了颠覆与创新，但在“大数据热”的浪潮推动下，我们应客观对待健康大数据发展尚不成熟这一事实，谨慎对待其分析结果。

参 考 文 献

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

[2] Jiang QP. The Change Force of Big Data [J/OL]. CIWEEK, 2013 (1) [2015-04-29]. http://mall.cnki.net/magazine/magadetail/HLZK201301.htm. (in Chinese)
姜奇平. 大数据的时代变革力量[J/OL]. 互联网周刊, 2013 (1) [2015-04-29]. http://mall.cnki.net/magazine/magadetail/HLZK201301.htm.

[3] Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications [J]. Health Affairs, 2014, 33(7): 1115-1122.

[4] Feng ST. Wearable devices development status and trend [J]. Inform Communic Technol, 2014(3): 52-57. (in Chinese)
封顺天. 可穿戴设备发展现状及趋势[J]. 信息通信技术, 2014 (3): 52-57.

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

[6] Fodeh S, Zeng Q. Special issue on mining big data in biomedicine and health care[J]. J Biomed Inform, 2014, 51: 1–2.
[7] Marx V. Biology: the big challenges of big data[J]. Nature, 2013, 498(7453): 255–260.
[8] Zou BJ. Big data analysis and its application in medical field [J]. Comp Educ, 2014(7): 24–29. (in Chinese)
邹北骥. 大数据分析及其在医疗领域中的应用[J]. 计算机教育, 2014(7): 24–29.
[9] Lv MT, Li DP, Wu J, et al. On the present condition of electronic health records and prospect [J]. Med Soc, 2006, 19(7): 60–61. (in Chinese)
吕孟涛, 李道苹, 吴静, 等. 电子健康档案现状分析与展望[J]. 医学与社会, 2006, 19(7): 60–61.
[10] National Heart, Lung, Blood Institute, Boston University. Framingham Hearth Study [EB/OL]. [2015–04–29]. http://www.framinghamheartstudy.org/.
[11] International Coordinating Centre, China Kadoorie Biobank, Nuffield Department of Population Health, University of Oxford. Kadoorie Study of Chronic Disease in China [EB/OL]. [2015–04–29]. http://www.ckbiobank.org/site/.
[12] van Den Eynden V. Maximisation of the value of population health sciences data[J]. The Lancet, 2012, 380(S3): S76.
[13] Clinicaltrials.gov [EB/OL]. [2015–04–29]. https://clinicaltrials.gov/.
[14] Young SD. Behavioral insights on big data: using social media for predicting biomedical outcomes[J]. Trends Microbiol, 2014, 22(11): 601–602.
[15] Topol E. The creative destruction of medicine: how the digital revolution will create better health care [M]. New York: Basic Books, 2012.
[16] Groves P, Kayyali B, Knott D, et al. The 'big data' revolution in healthcare[EB/OL]. [2015–04–29]. http://healthcare.mckinsey.com/big-data-revolution-us-healthcare.Published.
[17] Khoury MJ, Ioannidis JPA. Big data meets public health [J]. Science, 2014, 346(6213): 1054–1055.
[18] Gates B. The next epidemic-Lessons from ebola [J]. N Engl J Med, 2015, 372(15): 1381–1384.
[19] Fung IC, Tse ZT, Fu KW. Converting big data into public health [J]. Science, 2015, 347(6222): 620.
[20] Young SD. A "big data" approach to HIV epidemiology and prevention[J]. Prevent Med, 2015, 70: 17–18.
[21] Cohen J, Dolan B, Dunlap M, et al. MAD skills: new analysis practices for big data [J]. Proc VLDB Endow, 2009, 2 (2): 1481–1492.
[22] Schneeweiss S. Learning from big health care data[J]. N Engl J Med, 2014, 370(23): 2161–2163.
[23] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care[J]. Nat Rev Gen, 2012, 13(6): 395–405.
[24] Ghani KR, Zheng K, Wei JT, et al. Harnessing big data for health care and research: are urologists ready? [J]. European Urology, 2014, 66(6): 975–977.
[25] Kao RR, Haydon DT, Lycett SJ, et al. Supersize me: how whole-genome sequencing and big data are transforming epidemiology [J]. Trends Microbiol, 2014, 22(5): 282–291.
[26] Velthuis EJ, Malka ES, Richards MS. 'Big data' in health care. What does it mean and will it make a difference? [J]. Value Health, 2013, 16(7): A479.
[27] Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework[J]. J Gen Intern Med, 2013, 28(3): 660–665.
[28] Neubeck L, Lowres N, Benjamin EJ, et al. The mobile revolution-using smartphone apps to prevent cardiovascular disease[J]. Nat Rev Cardiol, 2015, 12(6): 350–360.
[29] Logan AG. Transforming hypertension management using mobile health technology for telemonitoring and self-care support [J]. Can J Cardiol, 2013, 29(5): 579–585.
[30] Toh S, Platt R. Is size the next big thing in epidemiology? [J]. Epidemiology, 2013, 24(3): 349–351.
[31] Howe D, Costanzo M, Fey P, et al. Big data: The future of biocuration[J]. Nature, 2008, 455(7209): 47–50.
[32] Murdoch TB, Detsky AS. The inevitable application of big data to health care[J]. JAMA, 2013, 309(13): 1351–1352.
[33] De Rosa M, Rossi E, Cataudella S. The role of big data in health care decision making: an italian experience [J]. Value Health, 2014, 17(3): A20.
[34] Zhang GL, Sun J, Chitkushev L, et al. Big data analytics in immunology: a knowledge-based approach[J]. Bio Med Res Int, 2014, 2014: 437987.
[35] Heinis T. Data analysis: approximation aids handling of big data [J]. Nature, 2014, 515(7526): 198.
[36] Psaty BM, Breckenridge AM. Mini-sentinel and regulatory science — Big data rendered fit and functional [J]. N Engl J Med, 2014, 370(23): 2165–2167.
[37] Lindenmayer DB, Likens GE. Analysis: don't do big-data science backwards[J]. Nature, 2013, 499(7458): 284.
[38] Chiolero A. Big data in epidemiology: too big to fail? [J]. Epidemiology, 2013, 24(6): 938–939.
[39] Ioannidis JPA. Why most published research findings are false [J]. Chance, 2005, 18(4): 40–47.
[40] Miller AR, Tucker C. Health information exchange, system size and information silos[J]. J Health Econom, 2014, 33: 28–42.
[41] Melnick G, Keeler E. The effects of multi-hospital systems on hospital prices[J]. J Health Econom, 2007, 26(2): 400–413.
[42] Lynch C. Big data: how do your data grow?[J]. Nature, 2008, 455(7209): 28–29.
[43] Khan N, Yaqoob I, Hashem IAT, et al. Big data: survey, technologies, opportunities, and challenges [J]. Sci World J, 2014, 2014: 712826.

(收稿日期: 2015–05–10)
(本文编辑: 王岚)