

大数据概念在医学影像中的应用探索

刘义, 王霄英

【摘要】 大数据(big data)概念的兴起是随着信息技术和统计技术的发展而来的,在商业和社会科学领域展现了一定的应用价值,在健康领域的应用也初步得到认可。但由于医学影像专业的特点,大数据的应用仍处于初步探索阶段。本文对大数据在医学影像学领域的应用进展作一综述。

【关键词】 大数据;影像信息学;医学影像学

【中图分类号】 R814.4 **【文献标识码】** A **【文章编号】** 1000-0313(2016)12-1124-03

DOI:10.13609/j.cnki.1000-0313.2016.12.003

“大数据(big data)”是近十年来被信息、商业领域广泛提及的一个概念,是近年来数据存储和计算能力、统计分析能力显著提高而衍生出的数据处理、获得利益的能力。大数据概念的兴起是随着信息技术和统计技术的发展而来的,在商业和社会科学领域展现了一定的应用价值,在健康领域的应用也初步受到认可。但由于医学影像专业的特点,大数据的应用仍处于初步探索阶段。

大数据基本概念

通常认为大数据是具有以下特征的数据集:体量大(volume)、种类多(variety)、生成速度快(velocity)、有潜在挖掘价值(veracity)。处理大数据超过传统数据分析技术的能力,需很强的存储、计算能力,数据统计和结果呈现也异于传统方法。实际上,现阶段对大数据的定义尚未得到公认,这个概念有很强的商业意味,在不同的应用场景下、对不同的人有不同的含义,大数据通常被用来描述和定义信息爆炸时代产生的海量数据,并命名与之相关的技术发展与创新。在大数据理念中,数据被认为是具有价值(价格)的资源,大数据的关键价值是新的应用,技术的关键要素是云技术和开源性分析工具(如人工智能、Hadoop^[1]和数据挖掘工具)等。

与传统科学研究的“数据”概念不同,大数据不追求精确度和因果关系,而是承认混杂性,探索相关关系,使用数据的直接目的是开发新产品和新服务。随着计算效率的指数级提升、数据的数量、深度和广度的增加、数据分析理论的进步以及现代社会对于个性化的追求,新的产业应运而生,全社会及每个人的决策等都会受到影响。无形的数据与有形的矿产、石油等资源一样,在合理利用的情况下,具有成为巨大经济资产的可能,带来全新的创业方向、商业模式和投资机会。

大数据在健康领域的应用

大数据在健康领域应用的第一个著名案例是 2009 年 Google 成功预测了甲型 H1N1 流感病毒的爆发^[2],在流感爆发前几周,研究论文已发表在《Nature》杂志上。研究论文不仅预测了全美国范围内将要传播的冬季流感,而且具体到特定的地

区和州。这项研究使用的预测依据与传统流行病学检测疾病的方式完全不同,并未使用任何医学检测手段(如血清化验检查、病原学检查等),而是分析了大众在网上搜索的词条。Google 把数千万条美国人最频繁检索的词条与近年来流感传播时期的数据进行了对比,处理了上亿个模型,挖掘出了特定检索词条的频繁使用与流感之间的联系。由于其庞大的数据资源、处理能力和统计技术,预测流感的结果与疾控中心数据的相关性达 97%。更重要的是,与传统流行病学数据的发布不同,这篇文章是在流感爆发前预测了疫情的发生,而传统的疾控数据是在疫情爆发之后一两周才能得到。由于 Google 保存了多年来所有的搜索记录,而且每天都会收到来自全球超过 30 亿条的搜索指令,正是由于庞大的数据资源以及相应的处理能力才能支撑它完成这项工作^[3]。这个大数据应用的案例引起了公共卫生专家的极大关注,此后,大数据在健康领域的应用逐渐增多,其优势逐渐被认识。

随着医学技术的进步,不同尺度的生物和临床数据以前所未有的速度和规模产生并被收集起来,用于健康管理^[4]。在微观领域,新一代基因测序技术每天可以处理数十亿的 DNA 序列数据;在宏观领域,广泛应用的电子健康病历(electric health record, EHR)可记录大量患者个体及人群的数据。面对迅速产生的海量数据,传统的方法不具备处理能力,而大数据为生命科学领域的研究者提供了从中挖掘规律的工具^[5]。以 IBM Watson 为代表的健康大数据分析工具已应用于药物研发和辅助制订临床决策^[6]。在患者就诊过程中,大数据分析可以帮助临床医师做出治疗决策^[7]。在住院过程中,大数据分析工具可以实时记录监护患者(如新生儿)的大量生命体征数据进行分析,及时发现患者的异常,以便针对性地采取措施,能明显改善患者的预后^[8]。出院后,大数据管理工具可以帮助患者监测健康状况,将患者健康信息反馈给管理中心,对患者的服药、就诊等进行提示^[9]。总之,在实际应用方面,大数据在健康领域快速渗透,在过去的几年里有许多新的发现和方法获得推广;在学科方面,大数据对以下几个生物医学领域的发展有重要促进作用:生物信息学、临床信息学、影像信息学和公共卫生信息学。

大数据在医学影像领域的应用探索

大数据可用于医学影像专业问题的解决和流程优化。医学影像大数据主要分两类:物理世界测量的数据(如图像和检

作者单位:100034 北京,北京大学第一医院医学影像科

作者简介:刘义(1989—),女,河北石家庄人,博士研究生,主要从事影像诊断研究工作。

通讯作者:王霄英, E-mail: cjr.wangxiaoying@vip.163.com

查过程信息)和人类活动产生的影像解读数据(如报告)。与其它领域的大数据比较,医学影像大数据的特点是客观性强、内容复杂、体量大、获取成本高和复用性强等。大数据分析的主要应用方向:1)实现个性化图像解读;2)促进研发新的成像技术;3)量化影像服务的价值;4)优化影像工作流程。

大数据在医学影像领域的最早应用是卫生经济学分析。医学影像检查在整体医疗费用里占了很大比例^[10-11],所以从降低费用、提升价值的角度引起了广泛的关注^[12-13],业界提出了“Value-based radiology”的概念^[14],影像的收费不应基于检查数量,而是基于影像信息对患者诊治的作用,且对患者和付费者公开透明^[15]。大数据研究分析了人群中大量影像检查的项目、费用和结果,提出影像检查对患者整体诊治结局、整体花费的影响,为影像检查适应证的制订提供帮助:在申请过程中,可基于数据分析提示申请者应选择何种检查方法及相关的费用情况^[16],对中风、肿瘤患者的合理化检查建议已获得较好的评价^[17-18]。市场中已有不少相关产品,典型的如美国放射学会(ACR)发布的 Appropriateness Criteria,根据大量循证医学证据和专家意见,对影像诊断、介入治疗的适应证给出了较全面的建议^[19-20]。

大数据在医学影像业务领域的应用也在广泛探索中。因影像检查已基本数字化,所以天然具有能进行大数据处理的可能。但在医学影像大数据的探索过程中,要解决三类问题。第一,影像检查的数据量极大,要求极高的存储、处理和分析能力,进行大数据研究的前提是有功能足够强大的硬件和软件的支撑,更需要有 IT 和统计专业人士的参与^[21]。第二,在数据利用过程中,简单的数据堆积不能进行有效的处理,需要按一定逻辑,从微观到宏观进行数据加工后方可挖掘其内在规律。在数据加工过程中,首先应从单个病例的影像中提取出关键信息(如美国国立卫生研究院的 common data element, CDE),实现图像中数据元素的标准化、结构化^[22],其次应开展高质量、大范围 and 长期的注册研究和临床试验^[23],并且将影像数据与临床数据、实验室检查、病理和基因组信息等整合在一起^[17],基于特定的假设和目标^[24-25]进行大数据分析,才能得到有意义的结果。第三,医学影像研究的思路应改变,传统的医学研究是先提出研究假设(如假设某种因果关系的存在),以临床观察或对照临床试验的形式验证假设或拒绝假设,研究数据通常是有目的收集,数据质量很高。但在大数据研究中,应允许数据有瑕疵,在不能确知因果关系之前,接受以相关关系来替代因果关系辅助决策,以整体数据预测个体结局^[26-27]。只有逐步适应这些研究思路的改变,医学影像研究者才能更好地利用大数据理念,并使其逐步完善。在完成实际任务的过程中,上述三类问题常常是互相叠加的,尚没有很简单的应对措施。所以总体上讲,大数据在医学影像业务领域的探索还是非常初步的,理论意义大于实际意义。

医学影像大数据工作的困难和解决方向

虽然当今医学影像的数据总量很大,但只有极小的一部分能被整合、理解和分析,面对海量数据研究的困难在于数据量极大、数据源过多、数据格式不统一以及瑕疵数据充斥数据库。在大数据处理过程中,统计理论和机器学习的技术非常重要,但医学影像专家对 IT 技术所知有限,而 IT 技术专家则不易理

解医学问题的实质,跨学科人才稀缺。

为了克服医学影像大数据工作的困难,应研发适应影像信息存储、处理和挖掘的规范、技术和工具。为了更好地利用医学影像大数据,应提高数据对使用者的透明度和方便性,提高数据使用效率和数据质量,对影像数据进行定量化、结构化地分析和挖掘^[28-29]。具体地说,应在保障患者隐私和数据安全的基础上,建立不依赖于厂家的存储和分析平台,使用统一的术语,发展定量的影像组学、影像共享、数据挖掘和人工智能工具等^[30]。

为了克服医学影像大数据工作的困难,应尽快培养跨学科人才,特别是领军人才。大数据工作是系统工程,从确定任务、采集数据、计算分析到决策应用,是一个很长的流水线。在这个流水线上的每一个环节,都存在严重的人才空缺。领导大数据工作的医学影像专家则应当努力成为理解、掌控整个流水线的综合性人才,并在医学影像工作团队中推广大数据的概念和方法。在商业环境中,大数据的推广应用有一个明确的过程,包括四个阶段:教育、探索、参与和执行^[31]。我们可以参考这个过程,在本部门中宣传大数据的相关理念,与 IT 技术人员合作,从解决单项明确任务起始,鼓励大数据的研究、开发和利用,进而广泛推广,使医学影像学科整体借助大数据技术得到充分的发展。

总之,医学影像大数据是资源,与实体物资(自然资源、人力资源)一样,具有价值和价格,大数据研究和应用将带来医学实践的改变。在医学影像发展历史上的多次技术跃进中,我们都是跟随者。此次以提升影像价值为导向的技术革新中,中国与世界前沿的距离并不大,我们完全有创新甚至局部领先的可能性。中国医学影像从业者应积极参与到技术革新中,有勇气去接受挑战,引领行业的发展。

参考文献:

- [1] White T. Hadoop: the definitive guide[M]. Sebastopol, CA: O'Reilly Media, Inc.; 2012; 15-30. <http://www.java1234.com/a/javabook/javaweb/2015/0513/4212.html>
- [2] Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009, 457(7232): 1012-1014.
- [3] Butler D. When Google got flu wrong[J]. Nature, 2013, 494(7436): 155-156.
- [4] Luo J, Wu M, Gopukumar D, et al. Big data application in biomedical research and health care: a literature review[J]. Biomed Inform Insights, 2016, 8(1): 1-10.
- [5] Hoyt RE, Snider D, Thompson C, et al. IBM Watson analytics: automating visualization, descriptive, and predictive statistics[J]. JMIR Public Health Surveill, 2016, 2(2): e157. DOI: 10.2196/publichealth.5810.
- [6] Chen Y, Elenee Argentinis JD, Weber G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research[J]. Clin Ther, 2016, 38(4): 688-701.
- [7] 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.
- [8] Spitzer AR, Ellsbury D, Clark RH. The pediatrix baby steps? data warehouse—a unique national resource for improving outcomes

- for neonates[J]. *Indian J Pediatr*, 2015, 82(1): 71-79.
- [9] Dimitrov DV. Medical internet of things and big data in healthcare [J]. *Healthc Inform Res*, 2016, 22(3): 156-163.
- [10] Duszak R, Berlin JW. Utilization management in radiology. Part 2: perspectives and future directions[J]. *J Am Coll Radiol*, 2012, 9(10): 700-703.
- [11] Duszak R, Berlin JW. Utilization management in radiology. Part 1: rationale, history, and current status[J]. *J Am Coll Radiol*, 2012, 9(10): 694-699.
- [12] Parker L, Levin DC, Frangos A, et al. Geographic variation in the utilization of noninvasive diagnostic imaging: national Medicare data 1998—2007[J]. *AJR*, 2010, 194(4): 1034-1039.
- [13] Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems 1996—2010 [J]. *JAMA*, 2012, 307(22): 2400-2409.
- [14] Sarwar A, Boland G, Monks A, et al. Metrics for radiologists in the era of value-based health care delivery[J]. *Radiographics*, 2015, 35(3): 866-876.
- [15] Durand DJ, Narayan AK, Rybicki FJ, et al. The health care value transparency movement and its implications for radiology[J]. *J Am Coll Radiol*, 2015, 12(1): 51-58.
- [16] Siström CL, Dreyer K, Weilburg JB, et al. Images of imaging: how to process and display imaging utilization for large populations [J]. *AJR*, 2015, 204(4): 405-420.
- [17] 吴静云, 张晓东, 王蕊, 等. 前列腺 MRI 影像生物样本库的建设探索[J]. *肿瘤影像学*, 2016, 25(2): 123-126.
- [18] Regge D, Mazzetti S, Giannini V, et al. Big data in oncologic imaging[J]. *Radiol Med*, 2016 Sep 13. DOI: 10. 1007/s11547-016-0687-5. [Epub ahead of print].
- [19] Bonow RO, Douglas PS, Buxton AE, et al. American college of cardiology foundation. ACCF/AHA methodology for the development of quality measures for cardiovascular technology: a report of the American college of cardiology foundation/American heart association task force on performance measures[J]. *Circulation*, 2011, 124(13): 1483-1502.
- [20] Patel MR, Spertus JA, Brindis RG, et al. American college of cardiology foundation. ACCF proposed method for evaluating the appropriateness of cardiovascular imaging[J]. *J Am Coll Cardiol*, 2005, 46(8): 1606-1613.
- [21] Boubela RN, Kalcher K, Huf W, et al. Big data approaches for the analysis of large-scale fMRI data using apache spark and GPU processing: a demonstration on resting-state fMRI data from the human connectome project [J]. *Front Neurosci*, 2016, 6(9): 492-?.
- [22] Saver JL, Warach S, Janis S, et al. Standardizing the structure of stroke clinical and epidemiologic research data: the National Institute of Neurological Disorders and Stroke (NINDS) Stroke Common Data Element (CDE) project[J]. *Stroke*, 2012, 43(4): 967-973.
- [23] Weiner MW, Veitch DP, Aisen PS, et al. The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception[J]. *Alzheimers Dement*, 2013, 9(5): 111-194.
- [24] Margolis R, Derr L, Dunn M, et al. The national institutes of Health's big data to knowledge (BD2K) initiative: capitalizing on biomedical big data[J]. *J Am Med Inform Assoc*, 2014, 21(6): 957-958.
- [25] Pettigrew RI. BRAIN initiative to transform human imaging[J/OL]. *Sci Transl Med*, 2014, 6(244): 244ed16. DOI: 10. 1126/scitranslmed. 3009695
- [26] Murdoch TB, Detsky AS. The inevitable application of big data to health care[J]. *JAMA*, 2013, 309(13): 1351-1352.
- [27] Badawi O, Brennan T, Celi LA, et al. Making big data useful for health care: a summary of the inaugural MIT critical data conference[J]. *JMIR Med Inform*, 2014, 2(2): 22. DOI: 10. 2196/medinform. 3447
- [28] 秦岫波, 王蕊, 高歌, 等. 前列腺多参数 MRI 报告进展: 基于第 2 版前列腺影像报告和数据库系统的结构化报告的构建[J]. *肿瘤影像学*, 2016, 25(2): 111-116.
- [29] 王可, 刘庆, 郭小超, 等. 肝癌影像报告进展: 基于 LI-RADS 的结构化报告[J]. *肝癌电子杂志*, 2016, 3(1): 26-31.
- [30] Zhao K, Wang C, Hu J, et al. Prostate cancer identification: quantitative analysis of T₂-weighted MR images based on a back propagation artificial neural network model[J]. *Sci China Life Sci*, 2015, 58(7): 666-673.
- [31] Kansagra AP, Yu JP, Chatterjee AR, et al. Big data and the future of radiology informatics[J]. *Acad Radiol*, 2016, 23(1): 30-42.

(2016-11-02, 修回日期: 2016-11-22)