科学通报 2015 年 第 60 卷 第 30 期: 2836 ~ 2844
评 述 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

大数据在肿瘤预后预测中的应用现状和前景

刘文扬, 金晶*

中国医学科学院/北京协和医学院肿瘤医院放疗科, 北京 100021
* 联系人, E-mail: jingjin1025@163.com

2015-06-18 收稿, 2015-07-14 接受, 2015-09-01 网络版发表

摘要 技术进步催生了大数据时代, 依靠先进的数据平台可将临床记录、医学影像、基因信息等不同形式的数据, 以及来自不同地区的数据, 迅速而有效地有机整合, 并进行及时的计算和分析. 这为肿瘤预后预测医学研究工作带来了前所未有的契机. 本文简述了肿瘤预后预测的研究现状, 回顾了大数据在医疗领域的相关研究成果, 旨在为推进大数据技术在肿瘤预后预测研究中的应用提供参考和新思路.

关键词
大数据
肿瘤
预后预测

2014年, 世界卫生组织发布了全球癌症报告, 预测全球癌症病例将呈现迅猛增长态势, 由2012年的1400万人, 逐年递增至2025年的1900万人[1]. 根据全国肿瘤登记估计, 2010年全国新发恶性肿瘤约309万例, 死亡病例约196万例[2]. 因此, 不仅当前, 而且在将来相当长一段时期, 肿瘤的防治工作都具有重大的社会意义, 凸显出巨大的科学价值. 肿瘤的预后预测则是其中的关键性研究领域. 卫生部门以此作为政策制定的依据, 而临床医师也将其作为医疗决策的参考. 对于患者而言, 生存期作为他们最为关心的预后指标, 将对他们的人生规划造成诸多重大影响, 当然, 生活质量也有其不可忽视的意义[3].

在传统的预后预测研究中, 应用最为广泛的统计学方法为Kaplan-Meier非参数预测模型和Cox回归模型等[3]. 近年来, 鉴于数据数量的增加和形式的复杂化, 人工神经网络(artificial neural networks, ANNs)也开始得到应用, 因为它具有更好的自适应性[4], 允许变量之间存在任意非线性的关系, 而非线性关系正是自然界中最普遍的现象. 尽管相关研究者对其"黑箱"问题也存在担忧[5,6], 但目前ANNs已经被食品药品监督管理局(Food and Drug Administration, FDA)批准用于宫颈癌的预后预测, 而且也应用于前列腺癌、结肠癌及乳腺癌等多种实体肿瘤, 实践显示其疗效预测准确性优于传统方法[7~9].

当前, 使用范围最为普遍的预测体系为TNM分期系统(T: tumor, 原发肿瘤范围; N: lymph node, 淋巴结转移情况; M: metastasis, 远处转移情况). 该系统主要基于肿瘤原发灶的侵犯范围(T)、淋巴结累及情况(N)以及远处转移情况(M)提供预测信息并支持治疗决策. 研究者针对该系统在不同肿瘤中的价值, 进行了广泛研究, 并尝试加入新的指标以便提高其准确性. 例如, 在口腔癌中, 研究者将肿瘤病理学特征与经典的预后因素进行结合从而提高了预测能力. 但TNM体系也不乏缺陷, 主要包括对基线参数分类不当、竞争性结果以及偏倚造成的预后低估或者高估. 尤其在涉及竞争性风险因素时偏倚明显, 如年龄, 如果将年龄作为肿瘤专项生存(cancer-specific survival, CSS)的相关因素考虑, 在临床研究中, 老年患者比例往往偏低, 从而导致该模型出现偏倚[3]. 近10年来, 基于分子肿瘤学的进步, 学者们也在各种肿瘤研究中尝试将预后基因指标加入模型, 如在肺癌研究中, 将个体化的分子基因信息与流行病学风险因素整合, 可用于确定适合进行后续治疗的患者. 但在发表的研究成果中, 大多数预后相关基因谱都很难重复[10].

引用格式: 刘文扬, 金晶. 大数据在肿瘤预后预测中的应用现状和前景. 科学通报, 2015, 60: 2836–2844
Liu W Y, Jin J. Big data in outcome prediction of cancer: Current landscape and perspective (in Chinese). Chin Sci Bull, 2015, 60: 2836–2844, doi: 10.1360/N972015-00161

目前最常用的循证医学体系中，随机对照研究(randomized clinical trial, RCT)仍然是级别最高的证据，但目前的部分RCT研究仍存在不少缺陷，无法满足复杂的临床需求. 首先，入组条件较为严格，通常能入组的样本仅占总体的20%左右，那么如何处理剩下80%的患者，尤其是那些合并症多，年龄大的患者. 其次，RCT往往在大型研究中心进行，其诊疗方案在普通医疗机构往往难以实施. 最后，某些诊疗方案由于伦理问题无法进入RCT，即使是公认有益的新技术，如调强放疗技术也难以进入RCT[11,12]. 最后，RCT往往耗时长，较观察性研究消耗的人力及物力更大[13~15]. 针对这些弊病，美国的研究者已利用大数据的方法大规模推进疗效比较研究(comparative effectiveness research)[16].

1 大数据的概念及对医疗行业的意义

大数据，即所谓巨量数据库. 这是一个新兴的概念，但目前尚无具体的界定标准. 例如，Victor和Kenneth[17]所描绘的大数据并非简单的数量巨大，而是对基于精确、因果、概率等传统观念的革命. 大数据拥有“4V”特征：(ⅰ) 容量巨大(volume)，大到令一般的软件已无法在合理的时间内完成处理；(ⅱ) 多样(variety)，数据结构和形式多样，并且无规则；(ⅲ) 高速(velocity)；(ⅳ) 价值(value). 大数据的大容量特点虽然给存储、处理及统计分析均带来了巨大挑战，但其丰富的数据资源也为相关研究带来了前所未有的机遇[18,19].

伴随着可得数据容量的剧增和处理能力的快速发展，大数据引发了超越技术层面变革. 在研究思路方面，不用随机分析方法(抽样调查)，而用所有数据进行分析；不再苛求精确函数关系，而重在找出相关性，寻求解决问题的方法[17]. 具体而言，传统的医疗相关研究大多基于已有临床经验、知识，提出假设，对数据进行分析和逻辑推理，从有限的取样数据中窥探影响预后的因素，进而发现因果关系，而大数据研究是则是凭借足够巨大的数据规模，对全数据进行统计性搜索、比较、聚类、归纳等分析，找出相关性，而非特定的因果关系，基于这些相关性，研究者对新的知识进行探索，并尝试进行预测. 因此，研究的趋势将从经验性较强的假设驱动向更加客观的数据驱动转变. 以肿瘤预后研究为例，由于医疗行业的特殊性，同时受制于经费、人力、患者意愿等多种因素的限制，以往大多数研究所纳入的病例数不过数百、数千，因此所采集的信息也比较有限，研究者只能基于先前设计的方案对非常有限的预后因素或者干预手段进行比较分析，其中大多数RCT研究均是对2种或者几种治疗方案进行比较，如经典的直肠癌术前放疗相关系列研究[20~23]. 而大数据则通过先进的半自动化数据采集方式，整合跨区域、多维度的海量医学数据，从数据出发，寻找规律，并利用连续积累和更新的数据，对新发现的预测模型进行验证，不断提高模型预测能力，改进治疗方式，提高疗效. 例如，美国临床肿瘤学会(American Society of Clinical Oncology，ASCO)初步实施的CancerLinQ项目，目前即已采集了177000例乳腺癌患者的临床、基因等综合信息，目前研究尚在进行. 这种全新的大数据研究方法，更加接近总体样本，必将对传统的TNM系统以及循证医学体系构成有力补充，并推动肿瘤治疗进入新的时代.

医疗行业历史悠久，积累了大量的数据. 以往这些资料大多以纸质形式保存，近年来已有快速向电子化格式过渡的趋势. 伴随着医疗成本控制、信息共享以及行政监管等多方面需求的上升和信息技术(information technology, IT)的不断进步，医疗资料的电子化将得到进一步的推广[24,25]. 以美国为例，2011年其医疗系统产生了150 EB(exabytes)的数据，照此速度发展，将很快达到ZB(zettabyte)数量级[24] 当前由于移动电话、无线设备等移动医疗设备的普及，不仅使医疗数据变得更加庞大，而且传输迅速、实时可得，传统的分析工具和方法已难以应付如此庞大而丰富、随时更新的数据资源，大数据科学进入医疗行业是行业发展的必然，也必将带来整个行业的变革[26,27]. 以生物医药为例，随着高通量计算、大数据[28,29]、个体化和人群模型[30,31]的发展，新的大数据平台整合了基因数据、临床信息和人群数据，开启了肿瘤等疾病的风险预测、个体化治疗和随访管理的新时代[31~33].

2 大数据运算系统及方法在医疗领域的应用

针对大数据在数据格式上往往存在结构化、半结构化和非结构化数据的状况，目前已开发出具备多线程平行计算能力的分布式构架程序，并逐步应用于生物医学结果预测领域[18,34,35]. 这些程序构架

的设计基于计算机群(computing clusters)而非超级计算机, 即分布式文件系统(distributed file system, DFS)[34,35]. 其中的成功代表为Google开发的Map-Reduce程序构架[35~37]. MapReduce适于大规模数据集(大于1 TB)的并行运算, 主要采用了函数式编程中的“Map(映射)”和“Reduce(归约)”思想, 它极大地方便了不会分布式并行编程的编程人员将自己的程序运行在分布式系统上(如Java, Ruby, Python等). 而MapReduce在生物医学领域应用最为广泛的程序则是Apache Software Foundation开发的Hadoop[38,39]. 在这些构架上, 多种算法的使用对有效处理数据十分必要, 如Anytime算法可较好地满足对处理时间的要求, 这类算法即使在完成之前中断, 也能对问题返回有效处理, 如果继续运行, 那么解决方案将会得到进一步改善. 如果数据迅速而不间断地产生, 那么可采用所谓“概念漂移”(concept drift)的分析方法进行处理[40~42].

新的IT构架在大数据研究中也扮演着重要角色, 其中云计算是在性能和成本上比较平衡的解决方案之一[43~45]. 在我国人口密度较低的内蒙锡林郭勒盟, 该方案已在人群疾病管理中发挥了较好的作用, 项目对当地291087人口的高血压和糖尿病实施了有效管理, 并且项目组通过平台侦测到了胆囊切除手术增多的情况, 综合地理信息发现是水受到了金属污染所致, 这体现了大数据平台从数据中发现新问题的优势[45]. 云计算类型较多, 可满足健康医疗领域的几乎所有需求. 不同平台的适用范围不同, 私有云(private clouds)仅供机构内部使用, 物理设备可置于机构内部或者第三方服务机构(这样可省去硬件建设). 交互云平台(community clouds)则适合多中心合作项目的使用, 不同机构共享云设备及信息, 有利于对项目的实时监督和结果分析. 而公共云平台(public clouds)则最大限度地利用了该技术的弹性和包容性, 可供专业人士、普通用户共同输入数据并分享信息[46,47].

令人遗憾的是, 目前这些大数据手段在医疗卫生领域仍未得到足够的应用[48], 尤其是肿瘤预后预测方面的研究在国内尚未见报道. 本文拟从数据的挖掘和采集、数据平台的建立等不同视角, 概述这些新方法在相关医疗领域的研究状况, 力图揭示其优势和局限, 为更好地利用和优化生物医学信息资源, 提高我国肿瘤预后预测研究水平提供参考和新思路.

3 大数据在肿瘤预后相关研究中的应用策略

3.1 加强对已有数据资源的利用

大数据最核心的技术是数据挖掘. 合理利用和整合已有的数据, 依靠研究方法的革新, 即有可能获得突破性的重大发现. 在高通量检测时代, 以美国国家癌症研究所(National Cancer Institute, NCI)为代表的研究机构已经积累了海量的医学生物学信息数据. 在对此类数据的研究中, Davoli等人[49]采用大数据研究策略, 以云计算为基础将现有三大基因数据库进行整合, 采用具有自动筛选特异序列和结构功能的PolyPhen-2预测算法[50], 建立了Tumor Suppressor and Oncogene (TUSON) Explorer程序, 基于超过8200种肿瘤配对标本的基因信息, 发现不同类型基因在染色体上的分布和致癌能力与癌症基因组中的非整倍体和拷贝数量变异的复杂模式具有相关性. 该研究组据此在全球首次提出了非整倍体是癌症的驱动者而非癌症的结果, 该结论合理解释了百年未解的非整倍体与癌症的关系问题. 借助于美国NCI主导的TCGA基因大数据平台项目, 研究者对295例胃癌患者的标本进行了成组体细胞拷贝数分析、全外显子序列分析、成组DNA甲基化程度分析、mRNA序列分析、miRNA序列分析、成组反相蛋白分析, 通过对所得出的大量生物信息学数据进行聚类分析, 首次提出了具有特定分子生物学特征的胃癌分子分型, 包括EBV感染型、以高突变率为特征的MSI型等4种类型[51]. 因此, 如何将这些已有肿瘤样本、基因蛋白质等组学信息, 甚至分子影像图像信息, 与临床数据通过大数据手段进行整合, 从而提高肿瘤预后模型的效能, 非常值得进一步研究.

此外, 随着互联网技术的普及, 海量的文献信息变得易于获取, 研究者通过大数据手段合理利用这些已有的知识和信息促进预后预测研究, 也具备一定前景. 由于以往的肿瘤预后预测基因谱重复性差, 同一种癌症中, 不同报道的预后基因谱有时竟无一重复, Venet等人[52]回顾发表的乳腺癌预后基因谱甚至发现其中60%并不优于随机产生的基因谱. 而研究者通过Google开发的PageRank算法, 将网络上已发布的基因信息进行整合, 成功提高了肿瘤预后预测基因谱的效能, 而且所得的基因谱往往能重复验证[53], 包括结直肠癌、乳腺癌和胰腺癌[54~58]. 通过系

统比较发现, 由于疗效更易受外在因素如年龄、一般状况等影响, 疗效预后预测比诊断和分型更加困难. 因此, 在肿瘤预后预测中, 最佳的模式是将基因信息与临床信息进行整合[53,57]. 美国新开展的ClinGen项目(http://www.clinicalgenome.org/), 即致力于将在临床上得到解释的基因变异汇总以便进一步研究.

在肿瘤的临床诊治和防控监测实践中, 已经和正在产生的信息数据非常值得进一步挖掘, 西方国家已对此广泛展开大数据研究. 一些成熟的数据平台已经投入使用, 如Multiparameter Intelligent Monitoring in Intensive Care database (MIMIC-2, 已更新到2.6版). 该数据库已纳入2001~2008年在Beth Israel Deaconess Medical Center (Boston, USA)诊治的超过30000例重症监护室(intensive care unit, ICU)患者, 我国研究者也利用该系统进行了一些10000例以上的样本研究, 得出了死亡率、输液需求量等关键问题的预测指标[59~65]. 我国从2006年也开始以年报形式发布全国肿瘤登记报告, 但利用大数据平台对数据进行肿瘤预后信息挖掘仍然缺乏[66~70]. 而且, 国际上仍有许多有待开发的数据宝藏: 例如, 美国各州郡的年度癌症报告, 丹麦有可追溯至1943年的国家癌症报告[27]. 相比之下, 美国已着手针对这些人群水平的数据进行一系列的探索[71], 例如, 莫非特癌症研究中心(Moffitt Cancer Center)的TCC(total cancer care)方案, 作为一个全新的肿瘤诊疗和研究整合模式, 它基于可以分析和处理多种类型、多层次的数据的实时信息平台, 将临床数据和肿瘤标本信息有机融合, 贯穿一名患者的终身, 以期提高治疗在多维度上的精确性[33,72]. 这一先进的数据平台模式, 值得学习借鉴.

3.2 前瞻性建设大数据分析平台

由于肿瘤预后过程复杂, 影响因素多, 且准确性要求高, 因而前瞻性建设大数据分析平台十分必要. Steinberg等人[73]基于36944例的人群数据, 采用大数据分析平台REFS(Reverse Engineering and Forward Simulation)成功预测了代谢综合征的发生风险, 模型在人群和个体水平上均可应用, 工作特征(receiver operating characteristic, ROC)/曲线下面积(area under curve, AUC)的范围为0.80~0.88, 优于传统的统计工具, 并且从采集到数据分析用时仅3个月, 时效性很强. 这显示了大数据平台可能成为最具成本效益的医疗策略性工具. 在肿瘤研究方面, ASCO正在开发名为CancerLinQ(http://www.asco.org/institute-quality/cancerlinq)的平台, 该平台可整合患者从临床到基因等综合信息, 并进行共享. 在前期研究阶段已经采集了177000例乳腺癌患者的数据, 预计到2015年中期, 将推进至其他实体瘤中并满负荷运转. ASCO主席Clifford Hudis认为该平台的运作将会发现临床研究中所缺少的信息[27]. 这与北卡罗莱纳州的癌症信息综合系统(integrated cancer information and surveillance system, ICISS)类似[74], 借助此类平台许多研究者已经基于人群水平在肿瘤防控及治疗方面取得了许多成果, 如肿瘤治疗变异的影响[75,76]、肿瘤就诊率的人群及地理差异[77]、公共卫生系统投入问题(提示公共卫生投入增长与包括癌症在内的可预防性死亡率下降相关)[78]、肿瘤相关卫生政策的收效[79]、以及肿瘤治疗副反应、生存率研究[80]. 这种综合信息平台的先进性定位决定了它不仅可用于肿瘤预后预测研究, 还可以根据研究者的需求开发新的分析方法, 拓展其研究潜力, 例如, 区分观察性研究之间存在的不同类型的偏倚[16], 患者异质性研究以及疗效比较[81,82]. 因此, 前瞻性大数据平台的建立可更加有效地进行数据采集、挖掘和结果分析, 促进包括肿瘤预后预测在内的各类研究工作[74].

在这些平台的建设中, 诚然, 先进的IT技术和设备作为载体是前提条件, 但是, 如何协同各方共享数据, 并且保证数据共享过程中的患者利益, 才是最为关键和困难的一步. 以美国为例, 首先, 1996年生效的医疗电子交换法案(Health Insurance Portability and Accountability Act/1996, Public Law 104-19, 国内一般直接简称HIPAA)为数据传输和共享提供了法律基础. 其次, 自2004年开始, 由布什总统提出10年内在全美建成实现卫生信息共享的EHR(electronic health records)系统. 该系统基于美国国家卫生信息网络, 电子化记录患者的综合健康信息, 可供患者在不同的医疗机构之间跨区域、跨平台进行实时传送和共享. 美国为了推进这一国家层面的项目, 出台了一系列政策: 设立区域性基金, 支持医疗信息系统的建立; 直接提供资金支持EHR相关项目, 并提供低利率贷款; 修改相关法律法规, 以利于实施EHR; 联邦医疗保险还对采用EHR的单位给予鼓励等. 目前已有约80%的医生和60%的医院应用这一系统. 最终, ASCO基于相关法律和数据网络基础, 凭借其全球领

先的学术地位和影响力, 在肿瘤领域主导了CancerLinQ项目, 并设立专门的基金予以财力支持. 该平台可从EHR系统中提取数据, 同时整合诊疗机构自身系统的数据. 这与以往的研究不同在于, 传统的研究只是收集特定的患者队列的指定信息, 而CancerLinQ则采集所有患者的完整信息, 包括人口学、预约情况、付费编码、门诊记录、病史、体格检查、家庭状况、社会关系、诊疗报告、手术记录、实验室检查结果、处方和用药信息等. 通过定期信息传送, 该平台可实时、迅速地对数据进行处理和分析. 该平台解决了患者信息传输共享过程中的隐私保护问题, 平台从各个端口对带有患者身份的数据进行收集和储存, 然后对数据进行标准化处理, 之后, 将对数据进行身份化处理, 并设立强大的防火墙, 当下游的使用者应用数据时则无法读取患者身份信息. 同时, 项目还制定了一系列的保护和监管措施以保证数据应用符合伦理[83]. 目前, 试运行阶段已有15家临床医疗机构开始准备向平台上传数据[84]. 而前述的ClinGen项目, 则由美国国立卫生研究院(National Institutes of Health, NIH)从国家层面主导, 采用与CancerLinQ类似的机制, 专注于推动基因等组学研究者和临床医师整合生物信息学和临床信息数据, 并促进学术研究中心、商业实验室、公立医院等不同机构之间的信息整合和共享[85].

3.3 借助社交媒体平台进行肿瘤预后研究

目前微信、Twitter等社交媒体使用广泛. 据人民网报道, 截至2014年2季度末, 微信活跃帐户已达4.38亿. 社交媒体中蕴含的"大数据"可间接提供有关使用者的行为、症状或者疾病状况等信息. 系统性回顾研究显示, 目前社交媒体相关医学研究有描述性研究62项, 干预性研究7项[86]. 其中比较著名的案例有利用Twitter用户自发产生的数据预测流感爆发[87,88], 与疾病预防控制中心(Center for Disease Control and Prevention, CDC)的传统方法相比, 该方法更加及时, 可监测疫情每周的变化, 准确性达85%. 不过, Google流感趋势预测工具(Google flu trends)所发表的结果与疾控中心的传统方法差异近2倍, 其有效性近期也受到了质疑[89]. Young[90]以社交媒体和移动技术为基础, 开发出了可预测HIV高危行为的模型, 为应用大数据预防HIV提供了创新性的工具. 因此, 将新兴媒体的信息与传统数据相结合而非对立, 将有助于提高预测分析的效力和可靠性[91]. 由于自发发布的信息数据噪音较多, 对准确性要求较高的肿瘤预后研究可能难以实施. 不过, Zaid等人[92]对生活在不同国家的罕见妇科肿瘤患者进行观察性研究, 通过Facebook填写标准化的问卷, 包括症状、生活质量和有无复发等信息, 结果显示该方案可行, 并且非常迅速. 在我国, 大部分研究机构可能短期内无法拥有大数据系统或平台, 而社交媒体作为一种可行的方案, 具备普及面广、便捷易用的优势, 而且成本低廉, 利用它开展肿瘤预后相关研究, 尤其是进行肿瘤事件随访、生活质量观察等工作, 对提高研究质量具有非常重要的现实意义.

4 大数据应用于肿瘤预后预测研究的挑战和展望

医疗行业从来不缺乏数据, 只是由于以往技术落后导致无法处理这些海量的信息. 大数据时代的来临让利用这些海量的数据造福肿瘤患者成为可能, 但是如何将噪音最小化、合理解读大数据产生的结果, 仍极具挑战[93].

大数据的强项在于寻找相关性, 而无法明确其有无意义. 1854年, "现代流行病学之父"John Snow在对霍乱病原体一无所知的情况下, 以当时的技术条件, 所要处理的流行病学信息无异于当今"海量"数据, 但他通过逐一记录感染家庭的位置, 提出了宽街(broad street)抽水泵是霍乱疫情源头的正确假说, 并在干预性试验中得到了验证. 如果用大数据技术处理该问题, 虽然系统可将位置信息整理工作在数小时内完成, 但技术本身却无法代替科学家提出有效的假说. 此外, 大数据也可能产生"大错误(big error)", 如前述的流感疫情过度高估[89]; 还可能得出虚假相关性, 如"用于生产蜂蜜的蜂群数量与因吸食大麻而被逮捕的青少年数量呈显著负相关"[93]. 因此, 对于肿瘤学家而言, 依靠大数据的方法找出相关性, 基于坚实的专业水平, 建立合理的假说, 将研究推进到假说验证和干预试验的层面, 才可能让大数据的进步转化为临床疗效的提高.

大数据本身是观察性数据, 无可避免地存在很多偏倚[93]. 然而, 肿瘤预后本身是一个非常复杂的过程, 受到肿瘤异质性、个体间差异、治疗决策、医疗条件、随访管理等多种因素的影响, 但由于预测结论对于临床决策相当重要, 医患各方对其准确性都

有着近乎苛刻的要求[3]. 从大数据技术的要求出发, 反观医疗行业, 整体尚存在两大不足: (ⅰ) 信息技术设备落后, 而且标准化程度差[94]; (ⅱ) 数据储存分散, 共享意识不佳. 大数据要成功应用于肿瘤预后等医疗研究, 对数据的准确性、安全性、信息处理能力和共享程度等诸多方面均提出了更高的要求, 因此从研究工作一开始就应将以下内容重视起来: (ⅰ) 数据采集质量(包括准确性、及时性和完整性); (ⅱ) 数据的标准化和整合(尽可能采用统一编码, 增大结构数据的比例, 为数据整合奠定基础); (ⅲ) 高水平的信息共享(打破医院间、地域上的隔阂, 建立数据共享的高水平平台和长期机制); (ⅳ) 提高数据计算能力(加强机器学习算法、分布式算法等新兴工具的应用); (ⅴ) 加强数据安全性(防止数据泄露, 加强存储管理和隐私保护等). 在这庞大的系统工程面前, 仅依靠医师、医院的努力, 恐怕远远不够. 政府相关机构、专业学会应发挥主导作用, 用宏观的视角去制定长远规划, 考虑如何通过行业标准的制定甚至立法, 对数据的采集以及共享的范围进行规范, 对患者隐私和权利加以保护. 大数据技术能否全面提升肿瘤预后等医学研究水平, 这些都是亟待解决的关键性问题. 我国目前尚无明晰的大数据战略规划顶层设计, 虽然从2010年开始实施卫生信息建设“3521工程”, 目前仍未建立起统一平台[95]. 美国早在2012年3月29日就由总统奥巴马宣布开启了“大数据研究和发展计划”, 从国家层面推动全方位的系统工作. 医疗方面, 美国国立研究中心NCI已启动两大大数据平台: 癌症成像存档数据共享服务平台(TCIA项目, http://www.cancerimagingarchive.net), 目前已存储约40万套医学图像; 癌症基因组图谱服务平台(TCGA项目, http://cancergenome.nih.gov/), 通过纳入大规模基因组测序信息, 加速对癌症分子基础的认识, 并已提出胃癌的分子分型[33,51,93].

综上, 在大数据为肿瘤预后预测等医学研究所带来的历史机遇面前, 需要政府主导、专业学会推动、医院间联合、研究者与患者参与等多方通力合作, 以传统研究方法为基础, 加强知识整合, 借助循证医学的研究原则, 制定合理而长远的规划, 才能在肿瘤预后研究等医疗事业上让大数据带来的变革转化为现实[19].

参考文献

1 Stewart B W, Wild C P. World Cancer Report 2014. Geneva: WHO Press, 2014
2 Chen W, Zheng R, Zhang S, et al. Report of cancer incidence and mortality in China, 2010. Ann Transl Med, 2014, 2: 61
3 Taktak A F G, Fisher A C. Outcome Prediction in Cancer. Amsterdam: Elsevier Science Publisers, 2007
4 Ripley R M, Harris A L, Tarassenko L. Non-linear survival analysis using neural networks. Stat Med, 2004, 23: 825–842
5 Biganzoli E M, Boracchi P, Ambrogi F, et al. Artificial neural network for the joint modelling of discrete cause-specific hazards. Artif Intell Med, 2006, 37: 119–130
6 Bishop J B, Szpalski M, Ananthraman S K, et al. Classification of low back pain from dynamic motion characteristics using an artificial neural network. Spine, 1997, 22: 2991–2998
7 Hu X, Cammann H, Meyer H A, et al. Artificial neural networks and prostate cancer—tools for diagnosis and management. Nat Rev Urol, 2013, 10: 174–182
8 Ahmed F E. Artificial neural networks for diagnosis and survival prediction in colon cancer. Mol Cancer, 2005, 4: 29
9 Burke H B. Artificial neural networks for cancer research: Outcome prediction. Semin Surg Oncol, 1994, 10: 73–79
10 Fridley B L, Koestler D C, Godwin A K. Individualizing care for ovarian cancer patients using big data. J Natl Cancer Inst, 2014, doi:10.1093/jnci/dju080
11 De Neve W, De Gersem W, Madani I. Rational use of intensity-modulated radiation therapy: The importance of clinical outcome. Semin Radiat Oncol, 2012, 22: 40–49
12 Veldeman L, Madani I, Hulstaert F, et al. Evidence behind use of intensity-modulated radiotherapy: A systematic review of comparative clinical studies. Lancet Oncol, 2008, 9: 367–375
13 Wang S D. Opportunities and challenges of clinical research in the big-data era: From RCT to BCT. J Thorac Dis, 2013, 5: 721–723
14 Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? Br Med J, 1998, 316: 201
15 Nallamothu B K, Hayward R A, Bates E R. Beyond the randomized clinical trial: The role of effectiveness studies in evaluating cardiovascular therapies. Circulation, 2008, 118: 1294–1303

16 Meyer A M, Wheeler S B, Weinberger M, et al. An overview of methods for comparative effectiveness research. Semin Radiat Oncol, 2014, 24: 5–13
17 Viktor M S, Kenneth C. Sheng Y Y, Zhou T, Translated. Big data: A Revolution That Will Transform How We Live, Work, and Think (in Chinese). Hangzhou: Zhejiang People's Publishing House, 2013 [维克托·迈尔-舍恩伯格, 肯尼思·库克耶, 著. 盛杨燕, 周涛, 译. 大数据时代: 生活、工作和思维的大变革. 杭州: 浙江人民出版社, 2013]
18 Rajaraman A, Ullman J. Mining of Massive Datasets. Cambridge: Cambridge University Press, 2012
19 Mohammed E A, Far B H, Naugler C. Applications of the MapReduce programming framework to clinical big data analysis: Current landscape and future trends. BioData Min, 2014, 7: 22
20 Kapiteijn E, Marijnen C A, Nagtegaal I D, et al. Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer. N Engl J Med, 2001, 345: 638–646
21 Bosset J F, Collette L, Calais G, et al. Chemotherapy with preoperative radiotherapy in rectal cancer. N Engl J Med, 2006, 355: 1114–1123
22 Gerard J P, Conroy T, Bonnetain F, et al. Preoperative radiotherapy with or without concurrent fluorouracil and leucovorin in T3-4 rectal cancers: Results of FFCD 9203. J Clin Oncol, 2006, 24: 4620–4625
23 Sauer R, Becker H, Hohenberger W, et al. Preoperative versus postoperative chemoradiotherapy for rectal cancer. N Engl J Med, 2004, 351: 1731–1740
24 Fernandes L, O'Connor M, Weaver V. Big data, bigger outcomes: Healthcare is embracing the big data movement, hoping to revolutionize HIM by distilling vast collection of data for specific analysis. J AHIMA, 2012, 83: 38–43
25 Big Data and Analytics Key to Accountable Care Success. IDC Health Insights, 2012
26 Marx V. Biology: The big challenges of big data. Nature, 2013, 498: 255–260
27 Savage N. Bioinformatics: Big data versus the big C. Nature, 2014, 509: S66–S67
28 Community cleverness required. Nature, 2008, 455: 1
29 Shaikh A R, Prabhu Das I, Vinson C A, et al. Cyberinfrastructure for consumer health. Am J Prev Med, 2011, 40: S91–S96
30 Murphy S, Patlak M. A Foundation for Evidence-Driven Practice: A Rapid Learning System for Cancer Care: Workshop Summary. Washington: National Academies Press, 2010
31 Butte A J, Shah N H. Computationally translating molecular discoveries into tools for medicine: Translational bioinformatics articles now featured in JAMIA. J Am Med Inform Assoc, 2011, 18: 352–353
32 Dalton W S, Sullivan D M, Yeatman T J, et al. The 2010 Health Care Reform Act: A potential opportunity to advance cancer research by taking cancer personally. Clin Cancer Res, 2010, 16: 5987–5996
33 Shaikh A R, Butte A J, Schully S D, et al. Collaborative biomedicine in the age of big data: The case of cancer. J Med Internet Res, 2014, 16: e101
34 Coulouris G, Dollimore J, Kindberg T. Distributed systems: Concepts and design. IEEE/ACM J Netw, 2005, 18: 182–231
35 de Oliveira Branco M. Distributed Data Management for Large Scale Applications. Southampton: University of Southampton, 2009
36 Zou Q, Li X B, Jiang W R, et al. Survey of MapReduce frame operation in bioinformatics. Brief Bioinform, 2014, 15: 637–647
37 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Commun ACM, 2008, 51: 7
38 O'Driscoll A, Daugelaite J, Sleator R D. "Big data", Hadoop and cloud computing in genomics. J Biomed Inform, 2013, 46: 774–781
39 Dong X, Bahroos N, Sadhu E, et al. Leverage hadoop framework for large scale clinical informatics applications. AMIA Jt Summits Transl Sci Proc, 2013, 2013: 53
40 Stella F, Amer Y. Continuous time Bayesian network classifiers. J Biomed Inform, 2012, 45: 1108–1119
41 Bellazzi R. Big data and biomedical informatics: A challenging opportunity. Yearb Med Inform, 2014, 9: 8–13
42 Wolfe P J. Making sense of big data. Proc Natl Acad Sci USA, 2013, 110: 18031–18032
43 Wall D P, Kudtarkar P, Fusaro V A, et al. Cloud computing for comparative genomics. BMC Bioinformatics, 2010, 11: 259
44 Kudtarkar P, Deluca T F, Fusaro V A, et al. Cost-effective cloud computing: A case study using the comparative genomics tool, roundup. Evol Bioinform Online, 2010, 6: 197–203
45 Lin C W, Abdul S S, Clinciu D L, et al. Empowering village doctors and enhancing rural healthcare using cloud computing in a rural area of mainland China. Comput Methods Programs Biomed, 2014, 113: 585–592
46 Kaur P D, Chana I. Cloud based intelligent system for delivering health care as a service. Comput Methods Programs Biomed, 2014, 113: 346–359
47 Zhou S, Liao R, Guan J. When cloud computing meets bioinformatics: A review. J Bioinform Comput Biol, 2013, 11: 1330002
48 Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Inf Sci Syst, 2014, 2: 3
49 Davoli T, Xu A W, Mengwasser K E, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell, 2013, 155: 948–962

50 Adzhubei I A, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods, 2010, 7: 248–249
51 Cancer Genome Atlas Research N. Comprehensive molecular characterization of gastric adenocarcinoma. Nature, 2014, 513: 202–209
52 Venet D, Dumont J E, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol, 2011, 7: e1002240
53 Roy J, Winter C, Isik Z, et al. Network information improves cancer outcome prediction. Brief Bioinform, 2014, 15: 612–625
54 van't Veer L J, Dai H, van de Vijver M J, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, 415: 530–536
55 Chuang H Y, Lee E, Liu Y T, et al. Network-based classification of breast cancer metastasis. Mol Syst Biol, 2007, 3: 140
56 Johannes M, Brase J C, Frohlich H, et al. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics, 2010, 26: 2136–2144
57 Winter C, Kristiansen G, Kersting S, et al. Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes. PLoS Comput Biol, 2012, 8: e1002511
58 Nibbe R K, Koyuturk M, Chance M R. An integrative-omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol, 2010, 6: e1000639
59 Saeed M, Villarroel M, Reisner A T, et al. Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database. Crit Care Med, 2011, 39: 952–960
60 Zhang Z, Xu X, Ni H, et al. Urine output on ICU entry is associated with hospital mortality in unselected critically ill patients. J Nephrol, 2014, 27: 65–71
61 Zhang Z, Xu X, Ni H, et al. Predictive value of ionized calcium in critically ill patients: An analysis of a large clinical database MIMIC II. PLoS One, 2014, 9: e95204
62 Saeed M, Lieu C, Raber G, et al. MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol, 2002, 29: 641–644
63 Mikhno A, Ennett C M. Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database. Conf Proc IEEE Eng Med Biol Soc, 2012, 2012: 5094–5097
64 Scott D J, Lee J, Silva I, et al. Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Med Inform Decis Mak, 2013, 13: 9
65 Lee J, Scott D J, Villarroel M, et al. Open-access MIMIC-II database for intensive care research. Conf Proc IEEE Eng Med Biol Soc, 2011, 2011: 8315–8318
66 Cooke C R, Iwashyna T J. Using existing data to address important clinical questions in critical care. Crit Care Med, 2013, 41: 886–896
67 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 Am Med Inform Assoc, 2014, 21: 957–958
68 Psaty B M, Breckenridge A M. Mini-sentinel and regulatory science—big data rendered fit and functional. N Engl J Med, 2014, 370: 2165–2167
69 Schneeweiss S. Learning from big health care data. N Engl J Med, 2014, 370: 2161–2163
70 Zhang Z, Xu X, Ni H, et al. Platelet indices are novel predictors of hospital mortality in intensive care unit patients. J Crit Care, 2014, 29: 885.e1–e6
71 Khoury M J, Gwinn M L, Glasgow R E, et al. A population approach to precision medicine. Am J Prev Med, 2012, 42: 639–645
72 Fenstermacher D A, Wenham R M, Rollison D E, et al. Implementing personalized medicine in a cancer center. Cancer J, 2011, 17: 528–536
73 Steinberg G B, Church B W, McCall C J, et al. Novel predictive models for metabolic syndrome risk: A "big data" analytic approach. Am J Manag Care, 2014, 20: e221–e228
74 Meyer A M, Olshan A F, Green L, et al. Big data for population-based cancer research: The integrated cancer information and surveillance system. N C Med J, 2014, 75: 265–269
75 Wheeler S B, Wu Y, Meyer A M, et al. Use and timeliness of radiation therapy after breast-conserving surgery in low-income women with early-stage breast cancer. Cancer Invest, 2012, 30: 258–267
76 Carpenter W R, Tyree S, Wu Y, et al. A surveillance system for monitoring, public reporting, and improving minority access to cancer clinical trials. Clin Trials, 2012, 9: 426–435
77 Holmes J A, Carpenter W R, Wu Y, et al. Impact of distance to a urologist on early diagnosis of prostate cancer among black and white patients. J Urol, 2012, 187: 883–888
78 Mays G P, Smith S A. Evidence links increases in public health spending to declines in preventable deaths. Health Aff (Millwood), 2011, 30: 1585–1593

79 Wheeler S B, Kohler R E, Goyal R K, et al. Is medical home enrollment associated with receipt of guideline-concordant follow-up care among low-income breast cancer survivors? Med Care, 2013, 51: 494–502
80 Kuo T M, Mobley L R, Anselin L. Geographic disparities in late-stage breast cancer diagnosis in California. Health Place, 2011, 17: 327–334
81 Methodology Committee of the Patient-Centered Outcomes Research I. Methodological standards and patient-centeredness in comparative effectiveness research: The PCORI perspective. JAMA, 2012, 307: 1636–1640
82 Luce B R, Kramer J M, Goodman S N, et al. Rethinking randomized clinical trials for comparative effectiveness research: The need for transformational change. Ann Intern Med, 2009, 151: 206–209
83 Schilsky R L, Michels D L, Kearbey A H, et al. Building a rapid learning health care system for oncology: The regulatory framework of CancerLinQ. J Clin Oncol, 2014, 32: 2373–2379
84 Masters G A, Krilov L, Bailey H H, et al. Clinical cancer advances 2015: Annual report on progress against cancer from the American Society of Clinical Oncology. J Clin Oncol, 2015, 33: 786–809
85 Rehm H L, Berg J S, Brooks L D, et al. ClinGen—the Clinical Genome Resource. N Engl J Med, 2015, 372: 2235–2242
86 Koskan A, Klasko L, Davis S N, et al. Use and taxonomy of social media in cancer-related research: A systematic review. Am J Public Health, 2014, 104: e20–e37
87 Chew C, Eysenbach G. Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One, 2010, 5: e14118
88 Broniatowski D A, Paul M J, Dredze M. National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic. PLoS One, 2013, 8: e83672
89 Lazer D, Kennedy R, King G, et al. Big data. The parable of Google Flu: Traps in big data analysis. Science, 2014, 343: 1203–1205
90 Young S D. A “big data” approach to HIV epidemiology and prevention. Prev Med, 2015, 70C: 17–18
91 Young S D. Behavioral insights on big data: Using social media for predicting biomedical outcomes. Trends Microbiol, 2014, 22: 601–602
92 Zaid T, Burzawa J, Basen-Engquist K, et al. Use of social media to conduct a cross-sectional epidemiologic and quality of life survey of patients with neuroendocrine carcinoma of the cervix: A feasibility study. Gynecol Oncol, 2014, 132: 149–153
93 Khoury M J, Ioannidis J P. Medicine. Big data meets public health. Science, 2014, 346: 1054–1055
94 Devaraj S, Ow T, Kohli R. Examining the impact of information technology and patient flow on healthcare performance: A Theory of Swift and Even Flow (TSEF) perspective. J Oper Manag, 2013, 31: 12
95 Han Z Y, Zhen T M, Gu J L, et al. 3521 project-based master design of regional health information construction framework. Chin J Med Libr Inform Sci, 2014, 23: 4

Big data in outcome prediction of cancer: Current landscape and perspective

LIU WenYang & JIN Jing

Department of Radiation Oncology, Cancer Institute (Hospital), Peking Union Medical College; Chinese Academy of Medical Sciences, Beijing 100021, China

In the era of big data driven by development of technology, different types of data, including clinical records, imaging, gene information, or even from different areas, can be combined effectively and promptly through the advanced informatics platforms. These platforms can also perform computation and data analysis. This progress provides unprecedented opportunity for medical research such as outcome prediction of cancer. This paper briefly describes the current landscape of outcome prediction in cancer, and reviews the relevant study results generated by big data approach in medical research, in order to promote the applying of big data technology in outcome prediction of cancer.

big data, cancer, outcome prediction

doi: 10.1360/N972015-00161