

• 综 述 •

肿瘤大数据在抗肿瘤药物应用中的研究进展

周海辉¹, 张海霞¹, 严婷婷², 葛卫红^{1*}

1. 南京大学医学院附属鼓楼医院 药学部 临床药理学室, 江苏 南京 210008

2. 南京大学医学院附属鼓楼医院 肿瘤中心, 江苏 南京 210008

摘 要: 世界范围内肿瘤大数据已经取得了长足的进步, 总结了肿瘤大数据在抗肿瘤药物应用中的研究进展, 包括肿瘤流行病学、肿瘤早期诊断、新靶点和新适应症的发现、不良反应的信号发掘、患者预后模型的构建, 以及肿瘤大数据存在的不足, 如数据的采集、可信度、保存、安全性等。

关键词: 肿瘤大数据; 抗肿瘤药物; 精准医学

中图分类号: R979.1

文献标志码: A

文章编号: 1674-5515(2018)11-3079-04

DOI: 10.7501/j.issn.1674-5515.2018.11.068

Research progress on tumor big data in the application of anti-tumor medicine

ZHOU Hai-hui¹, ZHANG Hai-xia¹, YAN Ting-ting², GE Wei-hong¹

1. Division of Clinical Pharmacy, Department of Pharmacy, Drum Tower Hospital Affiliated to Medical School of Nanjing University, Nanjing 210008, China

2. Cancer Center, Drum Tower Hospital Affiliated to Medical School of Nanjing University, Nanjing 210008, China

Abstract: Tumor big data has made great progress worldwide. The research progress of tumor big data in the application of tumor medicine including tumor epidemiology, early diagnosis, the discovery of new target and indications, signal discovery of adverse reactions, construction of patient prognosis model, and the shortage of tumor big data such as data collection, reliability, storage, and security, etc was summarized in this paper.

Key words: tumor big data; anticancer drugs; precision medicine

恶性肿瘤已成为威胁人类健康的重大疾病。世界范围内, 肿瘤治疗经历过传统的草药治疗、外科手术、合成化疗药物联合放疗、靶向治疗和免疫治疗。总体来说虽然取得了长足的进步, 但是在预防、治疗的各个环节依然存在很多问题。多种原因如多药耐药、基因多态性等使得医生的治疗手段很局限。在基因组学和生物信息学迅速发展, 尤以基因测序和互联网技术进步最为显著的今天, 肿瘤大数据和精准医学的结合将为肿瘤研究和治疗带来跨时代的变革。而肿瘤诊疗状况的权威数据关系到肿瘤防治、

临床研究、指南制定和政府决策。肿瘤大数据库可以提供癌症年度报告, 根据不同的分层预测肿瘤发病率、死亡率以及总体治疗获益; 癌症治疗与生存的年度报告, 包括肿瘤生存患者与诊断年龄分布和不同分期的治疗模式; 大数据基础上的肿瘤特征与预后分析。本文主要总结了肿瘤大数据在抗肿瘤药物应用中的研究进展包括肿瘤流行病学、肿瘤早期诊断、新靶点和新适应症的发现、不良反应的信号发掘、患者预后模型的构建, 以及肿瘤大数据存在的不足, 如数据的采集、可信度、保存、安全性等。

收稿日期: 2018-09-27

基金项目: 国家自然科学基金资助项目(81601153); 江苏省自然科学基金资助项目(BK20150104); 南京市医学科技发展基金资助项目(YKK16083); 南京市科技发展规划项目(201503020)

作者简介: 周海辉, 男, 主管药师, 博士, 主要从事临床合理用药及药理学研究。

*通信作者 葛卫红, 女, 主任药师。E-mail: gewhgy@126.com

1 肿瘤大数据对精准医疗的作用

1.1 促进肿瘤流行病学的发展

随着蛋白质组学等新兴学科的加入,肿瘤流行病学已经从当初的描述性研究、分析性研究发展到分子流行病学研究。1982年,PeIera和Weinstein首次提出“肿瘤分子流行病学”的概念,肿瘤分子流行病学是应用流行病学方法研究肿瘤标志物在高危人群和癌症患者中的分布和影响因素,对机体致癌物暴露、生物学效应以及个体遗传易感性进行测量和评价^[1]。通过对代表性人群从接触危险因素、癌前病变发展到肿瘤形成过程中一系列肿瘤标志物的研究,可以准确的测量“暴露”、生物学效应和遗传易感性。肿瘤标志物是连接实验室检测和传统流行病学研究的桥梁,通过对研究对象测定生物学标志物的定性或定量检测,可以评估致癌物暴露和机体遗传易感性的单独或联合作用。苯并芘-DNA加成物是人类发现的一个致癌物质的DNA加成物^[2]。20世纪70年代初,美国学者发现O⁶-烷基-2'脱氧鸟嘌呤加成物O⁶-MedG,并建立了单抗及放免技术^[3]。中国医学科学院利用这项技术进行了林县食管癌流行病学调查,找到了亚硝胺病因学说的证据^[4]。肿瘤大数据为肿瘤研究提供了新的重要手段和契机,大大增强了应用流行病学方法开展肿瘤危险性评价、筛查和诊治手段选择、预后评估、预防策略的制定能力。

1.2 肿瘤筛查、预防和早期诊断

遗传因素导致约10%肿瘤发生,获得特定人群肿瘤遗传易感性基因突变大数据,使对遗传性肿瘤的预防干预、早期诊断、精准分子分型诊断和精准的治疗成为可能,真正做到“上工治未病”。Oncomine库是全球最大的基因芯片数据库和整合数据挖掘平台,旨在发现能找到预测肿瘤发病的确定基因,针对性地开展肿瘤基因的筛查,并有助于发现新的治疗靶点和药物^[5]。

美国的肿瘤发病率为中国的1.6倍(高出65%),但5年生存率为中国的1.8倍(高出81%)^[6]。造成该差异的主要原因之一在于美国肿瘤早期筛查诊断的普及率更高,每年新发病例中仅24%患者为肿瘤晚期,而中国约50%新发病例为肿瘤晚期。2018年1月是宫颈癌宣传月,美国疾病防控中心的宣传语是任何女性都不应该死于宫颈癌。研究发现99.7%宫颈癌都是因感染人乳头瘤病毒(HPV)造成的^[7],默沙东研发的9价重组HPV疫苗早在2014前已经

在FDA获批上市,该疫苗通过预防HPV病毒感染,进而有效预防了宫颈癌的发病。2018年4月该疫苗在我国获批上市,大陆居民很快将会受益。

中国食管癌发病率高居世界首位,据估算,全球50%以上的食管癌患者都在我国^[8]。由于缺乏早期诊断的标志物,大部分患者发现时已是晚期,加上缺乏有效的临床治疗药物,患者的5年生存率仅为10%左右^[9]。另一方面,我国食管癌患者95%以上是食管鳞癌,而欧美国家多为食管腺癌,欧美的食管癌研究对我国医学界并无太多借鉴作用,进行食管癌全基因组的测序分析和功能研究,找出食管癌早期诊断的生物标志和药物靶标迫在眉睫^[10]。2015年,百度CEO李彦宏个人捐资3000万元,支持百度和协和医学院发起的针对食管癌基因的研究。此次合作研究将对1500位食管癌患者进行全基因组测序分析,在世界范围内,如此大规模的基因测序分析亦为首次。

1.3 发现肿瘤药物新的适应症

在临床诊疗实践中,每年都会有新的药物和诊疗手段出现。这些新技术要用于比临床研究中受试患者更广泛的人群,还面临着很多实际应用的问题。因为真实世界人群要远远大于药物、器械等研究环节的人群。2016年底,美国国会通过《21世纪治愈法案》,以及对可以利用“真实世界证据”取代传统临床试验进行扩大适应症的批准,更是确定了基于大量病例积累数据的真实世界研究的重要意义。国家明文规定药物上市以后5年内,制药企业必须要提交药物安全性监测数据,否则药物面临退市的风险。而基于真实世界的大数据则是制药企业完成这类法规需求的好方法。利用大数据挖掘和人工智能技术,从散布在世界各地的大量的专利、临床试验结果等海量信息中提取出能够推动药物研发的知识,提出新的可以被验证的假说,从而加速药物研发的过程^[11]。

1.4 不良发应(ADR)的信号挖掘

化疗是治疗恶性肿瘤的主要手段,但化疗药物对人体正常细胞缺乏选择性,在杀死肿瘤细胞的同时,也有严重的毒副作用。如抑制骨髓造血机能,胃肠功能紊乱等,很多患者甚至死于严重的骨髓抑制。有些不良反应(如骨髓抑制、胃肠道反应)较普遍,有些如(迟发性腹泻、肺纤维化、心脏毒性)只会出现在个别药物中。目前主流的从传统的自发呈报系统(SRS)获取不良反应的手段存在种种缺

陷,从其他途径获取又存在数据量小、获取难度大的问题。因此很多研究者都把目光投向医药信息大数据,针对性的获取自己感兴趣的学科和病种信息。目前很多医院都在利用医院 HIS 系统为 ADR 主动监测提供数据支持,其中包含大量的电子医疗记录(EMR)数据。潘雁等^[12]以 HIS 和临床检验信息系统(LIS)为基础,设计出化疗药物血液学不良反应监测软件,对常用肺癌化疗药物实施了血液学不良反应监测,为严重的不良反应提供预警作用。

1.5 新治疗靶点和药物的发现

传统的药物研发领域存在三大痛点:研发周期长、研发成本高、研发成功率低。医药大数据为新药研发开启了新篇章,对肿瘤患者的基因测序可以帮助了解癌症相关基因的变化,通过特定靶点研发新的靶向药物,如针对表皮生长因子受体(EGFR)突变阳性的非小细胞肺癌患者的奥希替尼、吉非替尼、厄洛替尼等。Atomwise 公司成立于 2012 年 6 月,公司总部位于美国旧金山硅谷,是一家利用超级计算机进行药品研发的公司。Atomwise 软件平台运行着 IBM 的蓝色基因/Q 超级计算机,分析已有数据库,并用人工智能和复杂的算法来模拟药品研发的过程,在研发的早期评估新药研发风险,并且该评估可以在几天内完成。

1.6 构建患者预后模型

传统的肿瘤预后研究多以单中心或多中心为单位,随访数据量有限、数据易缺失、可信度不够。临床急需数据量大、可信度高、更科学的预后评估体系。人工神经网络是模拟人脑的结构和信息传导方式建立的数学模型,已经在肿瘤预后研究中应用比较广泛,如时间编码增长模式和单时点模型^[13]。中国医科大学第一医院的 Shi 等^[14]利用 SEER 大数据库分析了 59 892 名甲状腺癌患者,利用 logistic 回归模型确定了影响老年患者预后的模型。英国学者 Enshaei 等^[15]构建了预测卵巢癌患者总生存率的人工智能模型,算法的准确性达到 93%,曲线下面积为 0.74,明显优于传统的统计工具。上述结果提示利用大数据分析构建肿瘤预测模型未来可能具有广阔的前景。

2 肿瘤大数据存在的问题

大数据通常有数据总量大、数据结构和类型复杂、单位数据价值密度较低、生成速度快等几个共同特征。因此对开展和从事大数据研究的医疗机构和相关专业人员提出了独特的需求和挑战。

2.1 数据的采集与整理

目前除了临床诊疗中被动积累的业务系统运行数据、电子病历和影像检查数据以外,医疗机构主动开展的肿瘤相关基础和临床研究也产生了具有庞大数据量或涉及大规模人群的基因组学、人群或队列筛查体检数据等,已成为肿瘤大数据的重要来源。但由于各自数据类型和结构特征的巨大差异,很少有成熟的医疗信息化平台或工具可以将之整合,导致付出很高成本获得的数据资源难以充分发挥其作用和价值。此外,尽管临床大数据的采集不再设置很高门槛,然而循证医学临床研究范式中的选择性偏倚、幸存者偏倚、测量偏倚等混杂偏倚和数据间的交互作用仍是设计数据分析策略、报表、发布研究结论等过程中需要充分考虑的问题,即肿瘤大数据研究本质上仍然属于医学研究,其结论的得出不应因为采用大数据技术而对方法的合理性或结论的可信度降低要求。只有把海量的数据进行科学的整合才能做到有效合理的利用。

2.2 数据的可信度

数据讲究准确性、完整性、时效性和一致性。由于部分医护人员的责任心不强,如套用病例模板、患者的个人信息收集不完整、未及时记录患者病情的变化等都会影响数据的可靠性。没有关于家族史、既往史、并发症及手术切缘状况、病理学资料的记载,没有局部复发情况的记录。在治疗方面的记录不完全。除了有手术及放疗的记录外,无辅助治疗、内分泌治疗、化疗、生物治疗等记录,缺少第一疗程的后续治疗相关记载。没有记载乳酸脱氢酶、淋巴结转移数目、全身疾病、血红蛋白水平等重要的预后因素。

2.3 数据的保存

目前各个医院基本都有自己的服务器或第三方机构的数据存储设备,但是单个的“信息孤岛”往往不能承载大量的数据。缺乏全国肿瘤诊疗状况的权威数据,这就会导致肿瘤防治、临床研究、指南制定、政府决策等带有一定盲目性。早在 20 世纪八九十年代,美欧日等发达国家率先建立三大生物数据中心:美国国家生物技术信息中心(NCBI)、日本 DNA 数据库(DDBJ)、欧洲生物信息研究所(EBI)^[16]。反观国内,我国 2017 年才成立了中国卫生信息学会健康医疗大数据肿瘤专业委员会暨慢病防治联盟指导委员会,填补了以医院为基础的癌症监查网络的空白。

2.4 数据安全性

信息化程度的提高为大数据的收集和存储提供了便捷,但同时也面临着诸多挑战。数据安全遭到诸多威胁,病毒和木马的入侵容易导致信息的丢失和泄露。为此,2016年国务院办公厅印发《关于促进和规范健康医疗大数据应用发展的指导意见》提出,加强健康医疗数据安全保障。安全和发展是相辅相成的,安全是发展的前提,发展是安全的保障。健康医疗大数据的应用发展,最基础的在安全,最重要的在发展。个人的健康医疗信息最为敏感,属于隐私保护范围,要依法进行严格管控保护,绝不能公开或泄露,一定加强应用安全风险评估和防范。大数据资料对个人数据的运用,有一个比较惯常的做法,就是“脱敏”“去标识化”,对个人隐私产生影响的这些内容一定要去掉,才能就某一种疾病进行大数据的挖掘分析。对涉及健康医疗数据的整个管理要有一套严格的法律法规。主要是建立数据安全管理制度,注重内容安全和技术安全。

3 结语

随着大数据和互联网技术的飞速发展,以及临床大数据研究开发的快速推进,肿瘤大数据研究未来的发展方向仍会继续坚持科学审慎原则,整合生物-医学大数据,推动临床真实需求导向和应用驱动的增量改进。重视医疗大数据的采集、存储、整合,提高准确度和安全性,让医疗大数据更好的造福全人类!

参考文献

- [1] Perera F P. Molecular epidemiology: on the path of prevention [J]. *J Natl Cancer Inst*, 2000, 92(8): 602-612.
- [2] Starostenko L V, Rechkunova N I, Lebedeva N A, et al. Human DNA polymerases catalyze lesion bypass across benzo[a]pyrene-derived DNA adduct clustered with an abasic site [J]. *DNA Repair (Amst)*, 2014, 24: 1-9.
- [3] Gottschalg E, Scott G B, Burns P A, et al. Potassium diazoacetate-induced p53 mutations in vitro in relation to formation of O⁶-carboxymethyl- and O⁶-methyl-2'-deoxyguanosine DNA adducts: relevance for gastrointestinal cancer [J]. *Carcinogenesis*, 2007, 28(2): 356-362.
- [4] 林东昕, 谭文, 陆士新, 等. 中国食管癌分析流行病学研究 [J]. *中华流行病学杂志*, 2003, 24(10): 939-943.
- [5] Rhodes D R, Yu J, Shanker K, et al. Oncomine: a cancer microarray database and integrated data-mining platform [J]. *Neoplasia*, 2004, 6(1): 1-6.
- [6] Rahib L, Smith B D, Aizenberg R, et al. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States [J]. *Cancer Res*, 2014, 74(11): 2913-2921.
- [7] Goodman A. HPV testing as a screen for cervical cancer [J]. *BMJ*, 2015, 350: h2372.
- [8] 贺宇彤, 李道娟, 梁迪, 等. 2013年中国食管癌发病和死亡统计 [J]. *中华肿瘤杂志*, 2017, 39(4): 315-320.
- [9] Thein H H, Jembere N, Thavorn K, et al. Estimates and predictors of health care costs of esophageal adenocarcinoma: a population-based cohort study [J]. *BMC Cancer*, 2018, 18(1): 694.
- [10] Lin Y, Totsuka Y, He Y, et al. Epidemiology of esophageal cancer in Japan and China [J]. *J Epidemiol*, 2013, 23(4): 233-242.
- [11] Williams A M, Liu Y, Regner K R, et al. Artificial intelligence, physiological genomics, and precision medicine [J]. *Physiol Genomics*, 2018, 50(4): 237-243.
- [12] 潘雁, 许海静, 朱珺. 化疗药物血液学不良反应监测软件的设计及应用 [J]. *中国药房*, 2011, 22(1): 87-89.
- [13] Ashbeck E L, Bell M L. Single time point comparisons in longitudinal randomized controlled trials: power and bias in the presence of missing data [J]. *BMC Med Res Methodol*, 2016, 16(1): 43.
- [14] Shi L Y, Liu J, Yu L J, et al. Clinic-pathologic features and prognostic analysis of thyroid cancer in the older adult: a SEER based study [J]. *J Cancer*, 2018, 9(15): 2744-2750.
- [15] Enshaei A, Robson C N, Edmondson R J. Artificial intelligence systems as prognostic and predictive tools in ovarian cancer [J]. *Ann Surg Oncol*, 2015, 22(12): 3970-3975.
- [16] Kosuge T, Mashima J, Kodama Y, et al. DDBJ progress report: a new submission system for leading to a correct annotation [J]. *Nucleic Acids Res*, 2014, 42(Database issue): D44-D49.