大数据对肿瘤登记发展的影响

魏矿荣^①, 刘慎超^②, 魏东霖^③, 梁智恒^①, 陈万青^{④*}

① 中山大学附属中山医院肿瘤研究所, 中山 528403;

② Department of Technology, Dimensional Insight, Massachusetts 01803, USA;

③ Department of Computer Science, Dartmouth College, New Hampshire 03755, USA;

④ 全国肿瘤防治研究办公室, 全国肿瘤登记中心, 北京 100021

* 联系人, E-mail: chenwq@cicams.ac.cn

2014-08-22 收稿, 2014-10-21 接受

摘要 对肿瘤登记和大数据的发展、大数据对肿瘤登记的作用和影响进行了系统回顾, 大数据和肿瘤登记近年发展快速, 大数据的发展促进了肿瘤登记的发展, 而肿瘤登记发展又推动了大数据的进步. 在大数据的影响下, 肿瘤登记将向以下几个方向发展: (i) 肿瘤登记自动化; (ii) 及时收集、整合和更新不同来源的肿瘤数据, 提高肿瘤资料的质量、可用性和易用性, 推动肿瘤资料的开放和共享, 扩展肿瘤资料的应用; (iii) 各级医疗卫生信息中心的出现将彻底改变肿瘤登记模式; (iv) 采用大数据的技术和方法建立各级肿瘤数据中心; (v) 医院肿瘤登记逐步开展, 人群肿瘤登记覆盖率大幅提高, 部分省将全民开展肿瘤登记. 肿瘤登记自动化的理念和技术, 迎合了大数据和肿瘤登记发展的趋势, 并可借鉴应用到其他疾病的监测和研究.

关键词

大数据
肿瘤登记
流行病学

恶性肿瘤对人类生命健康和社会发展危害极大, 据全球癌症报告(GLOBOCAN)估计, 2012年全球恶性肿瘤(除皮肤非黑色素瘤)发病和死亡数分别为1409万和820万例, 2015年将分别为1523万和888万例^[1], 恶性肿瘤死亡占2011年全球死因第1位^[2], 所以恶性肿瘤防治非常重要. 而有效降低恶性肿瘤的危害, 首先必须及时、完整、准确地掌握恶性肿瘤流行病学资料. 但是, 由于肿瘤资料数量巨大、种类多、分布广, 有关数据库相互孤立, 大部分为非结构化数据等原因, 目前传统数据收集方法已难以及时、完整和准确地收集肿瘤资料, 尤其是有关临床分期、治疗和随访等信息, 同时传统数据库管理办法也难以管理越来越多的、复杂的肿瘤数据, 严重制约了肿瘤登记工作的发展. 大数据技术为肿瘤登记和研究带来了便利, 采用大数据技术的方法, 可有效解决肿瘤登记和研究面临的困局.

虽然大数据一词由来已久, 但直到2011年才被

广泛关注, 中国知网中以“大数据”为关键字的论文, 2010年底前仅有15篇, 2011年后显著增多, 但有关大数据实际应用的文献较少^[3], 目前尚未查到关于肿瘤登记的文章. 因此, 为更好地开展肿瘤登记研究, 遏制恶性肿瘤危害性, 本文系统回顾了大数据和肿瘤登记的发展及其在研究领域的应用.

1 肿瘤登记的发展

1.1 人群肿瘤登记在肿瘤防治中的作用日益重要

肿瘤登记在肿瘤防治中的作用非常重要, 是肿瘤防治的基础. 联合国及其卫生机构非常重视提高国家和区域水平的癌症监测. 2011年联合国有关非传染性疾病高级会议的政治宣言重申了肿瘤监测的重要性, 2012年第65届世界卫生大会同意建立全球非传染性疾病防控监测框架. 国际肿瘤登记协会 (International Association of Cancer Registries, IACR)和国

引用格式: 魏矿荣, 刘慎超, 魏东霖, 等. 大数据对肿瘤登记发展的影响. 科学通报, 2015, 60: 491-498

Wei K R, Liu S C, Wei D L, et al. Influence of "big data" on the development of cancer registration methods (in Chinese). Chin Sci Bull, 2015, 60: 491-498, doi: 10.1360/N972014-00882

际癌症研究所 (International Agency for Research on Cancer, IARC) 极力推动全球尤其是中低收入国家, 如非洲和亚洲人群肿瘤登记的发展^[4]. 由于资源有限, 高收入和中低收入国家之间的癌症登记差异很大, 2007年北美地区肿瘤登记人群覆盖率达83%, 欧洲为32%, 拉丁美洲和亚洲分别为6%和4%, 而非洲仅为1%, 迫切需要增加非洲、亚洲和拉丁美洲人群覆盖率^[5]. 2011年11月世界癌症领导者高峰会议启动了GICR (Global Initiative for Cancer Registry Development) 项目, 内容包括建立区域癌症登记中心、提供技术支持、研究合作、培训和咨询^[4]等. 此外, IARC制定了新的技术规范, 扩大了肿瘤登记内容^[2], 积极推动肿瘤生存研究^[6]. 许多国家也积极响应, 努力推动人群肿瘤登记的发展, 如2013年不丹开始了人群肿瘤登记^[7]. 随着肿瘤登记的发展, IARC和各国充分认识到, 借助医疗、卫生、经济和卫生信息技术的发展, 尤其是电子数据库的完善, 肿瘤登记也应发生相应变化. 基于电子数据库和信息技术的新的登记方法不断涌现, 半自动和自动登记方法渐露头角, 大大方便和推动了肿瘤登记的开展^[8-11]. 2014年在加拿大渥太华召开的第36届国际肿瘤登记年会(36th IACR Conference 2014), 有专门关于“Methods of enhancing registration using electronic methods”的主题, 来自5个国家的演讲者分享了包括“Automatic data capture and coding systems”等在内的肿瘤登记自动化相关研究进展, 中国“自动查找和编码肿瘤数据”的演讲引起了广泛重视^[12].

中国自1959年河南林县开展肿瘤登记以来^[13], 肿瘤登记点数目不断增加, 覆盖人群不断扩大, 尤其是2008年获得中央财政转移支付资金支持后, 人群肿瘤登记快速发展^[14], 至2013年底中国肿瘤登记点数目达222个, 覆盖中国总人口的15.42%^[15], 2014年增加到308个, 预计人口覆盖将达到3亿人. 登记质量不断提高, 符合质量要求的资料才能被纳入年报^[15]. 肿瘤登记内容也不断增加, 2014年开始了肿瘤病例随访资料的上报和相关生存分析. 目前, 中国肿瘤登记逐步规范和标准化, 每年出版《中国肿瘤登记年报》, 每5年出版《中国癌症发病与死亡报告》, 并适时发布癌情报告供有关部门参考. 经过不懈努力, 目前中国肿瘤登记达到了世界中高水平^[2], 最新发布的《五大洲癌症发病》第10卷收录了中国12个登记处的资料^[16].

1.2 医院肿瘤登记的作用渐受关注

医院肿瘤登记不仅是人群登记的基础, 对医院管理、临床疗效评判和科研开展也非常重要, 因而渐受关注. 虽然国家尚未要求开展医院肿瘤登记, 但一些单位已充分认识到医院肿瘤登记的重要性, 正在或计划开展医院肿瘤登记. 河北省肿瘤医院除了负责全省的人群肿瘤登记, 也开始了医院肿瘤登记. 海南医学院第一附属医院为了提高医院临床和科研水平, 正筹划开展医院肿瘤登记, 并希望将病理组织标本、肿瘤临床和随访资料有机结合起来, 以更好地为医院管理、临床和科研服务. 中南大学启动了“湘雅临床大数据建设”项目, 成为我国首个开展大数据在临床医学领域系统应用的高校.

随着相关学科和技术的发展、肿瘤防治的需要和肿瘤登记的蓬勃发展, 肿瘤登记也面临一系列的问题与难题, 如数据收集、整理、储存、统计分析、共享、安全和隐私等问题, 传统的数据库管理方法逐渐难以胜任, 需要采用新的如大数据方法进行应对.

2 大数据的特点、应用及发展

大数据最初指大小超出了传统数据库收集、存储、管理和分析能力的数据集, 近年其含义明显扩大, 不仅指数据本身, 也包括一整套用于收集、存储、管理、分析大型数据和解决复杂数据问题的技术. 大数据被赋予各种含义, 表明信息技术发展已由“计算”转向“数据”, 大数据很好地概括了当前数据管理中的重要性和多样性. 大数据既是社会经济高度发展的结果, 也是信息技术发展的必然^[17].

2.1 大数据的特点

大数据无严格定义, 普遍认为大数据是指规模大而复杂、以至很难用现有数据库管理工具处理的数据集. 维基百科的定义是指采用常用软件工具处理所需时间超过可容忍时间的数据集^[18], 国际数据公司(International Data Corporation, IDC)认为, 大数据技术是一种新技术和架构, 这种技术和架构能很经济和高速度地提取超大规模数据中的价值. 大数据具有4“V”特点, 分别是大量(Volume)、多样性(Variety)、速度(Velocity)和精确性(Veracity). IDC认为, 大数据的第4“V”指的是数据的价值性(Value), 强调海量数据中包含重要价值, 也有人认为第4“V”指的是数据

的低密度价值性,即大数据中有价值的信息可能非常有限^[19]。

根据来源的不同,大数据可分为来自于人、机(计算机信息系统)和物(各类数字设备)3大类^[20],又可分成Web、决策和科学数据3大类。Web数据是与Web相关的数据,决策数据主要指商务智能数据,而科学数据是指科学活动产生的数据^[19],肿瘤数据主要是科学数据。数据产生大致经历了被动、主动和自动3个阶段,自动产生的数据是大数据产生的最根本原因,正是数据产生方式的巨变诞生了大数据^[19]。

大数据数量巨大、内容丰富、覆盖范围广,包含语种多、动态性与交互性并存、结构复杂,包含大量非结构或半结构化数据^[21],但其分析算法简单、速度快、数据处理量大,处理的样本是数据总量,过程不涉及原有业务,存储简单,以挖掘发现数据的潜在价值及其之间的关系为目的^[22]。大数据处理流程基本一致,包括数据抽取、集成、数据分析和解释等^[17,23]。不同领域大数据的应用有不同特点,而大数据的广泛存在使得大数据应用具有挑战性^[19],并创造了数据科学家这一热门职业^[3]。大数据思维是一种创新性思维,大数据时代是由数据开放引发的数据、技术和思维的变革^[24],第4种范式的实质就是数据思维,是以计算为中心到以数据为中心的思维转变^[19]。大数据的出现使传统数据库面临着许多挑战,在秒级时间内得出分析结果正是大数据和传统数据处理技术的最大区别^[25]。

大数据与云计算关系密切,云计算是大数据的基础和运行管理平台,使大数据应用成为可能,而大数据是云计算的具体应用,两者在技术和应用上相辅相成,云计算遭遇大数据,是时代发展的必然趋势。

2.2 大数据的应用

信息时代的竞争是知识的竞争,而知识的竞争主要是数据的竞争,数据竞争将成为经济发展的必然前提^[22]。预计全球2014年大数据投资近300亿美元,而未来6年投资复合增长率将达每年17%^[26]。

(i) 大数据在政府的应用。政府是最大数据拥有者,但其数据价值大部分以潜在形式存在,需要通过创新性应用来挖掘。同时政府数据利用率低,因而应开放政府数据,以提升政府数据的价值。2011年美国首先与印度政府合作开放数据,共同将美国

政府数据开放平台(data.gov)打造成开源平台。英国政府提出自由数据(free our data)和数据权(right to data)概念,倡导开放政府数据,并认为数据权是公民基本权利。澳大利亚的《公共服务大数据战略》注重大数据在公共服务中的应用,政策更具体细致、操作性更强^[27]。41个国家于2011年召开了开放政府数据国际研讨会,8个国家共同成立了开放政府联盟(OGP),并发布了《开放政府宣言》,全球已有31个国家和地区建立了数据开放网站^[23]。

(ii) 大数据在企业的应用。大数据成功应用于企业,中石油、中石化等大型企业以及各大银行拥有自己的数据中心,而智能电网、智能交通、智能医疗保健等发展也依赖于大型数据中心和新型商务智能能力的建设^[17]。中国工商银行的“大型商业银行基于数据仓库的精准营销管理”项目,标志着我国商业银行在应用大数据创新市场营销模式、提升企业现代化管理水平方面取得了新进展,大数据从概念化走向价值化^[28]。数据最优先的企业经营管理模式业已出现,大数据将催生大数据产业^[24],数据资源服务商将会不断出现^[23,29]。如运用得当,美国医疗服务的大数据每年可产生3千亿美元的效益^[30]。

(iii) 大数据应用的问题。大数据应用问题主要表现在数据收集、整合、分析、开放和共享、管理、安全和隐私保护、质量、可信、可用和易用性等,如何及时、完整、准确和便利地收集所需信息,是数据收集面对的主要问题^[19],整合来源和结构不同的数据是数据分析利用前提^[31]。大数据时代核心竞争力依赖于对大数据的分析挖掘能力,数据再利用、重组、扩展、开放、合作、试点推广、加强技能教育等可提高数据分析和价值发现能力^[23,24]。数据必须被应用才有价值,因而数据应该开放共享,其基本原则是完整、原始、及时、可读取、机器可处理、数据获取无歧视和无需许可、数据格式通用,科学数据共享已成为潮流^[17,24,31]。大数据管理规范化和标准化是大数据利用前提,目的是保证数据的完整和准确性,是确保大数据战略顺利实施的关键^[27]。大数据面临许多安全风险,如数据采集、共享、分析和发布时的安全与隐私保护,数据生命周期的安全和隐私保护,隐私数据可信销毁等,这些大数据安全和隐私问题^[20]应予以重视。数据质量是数据价值的保证,直接影响到数据可信和可用性,数据质量问题贯穿数据收集、使用和发布的整个过程,但常发生在数据发布时,应

制定和遵循数据质量原则^[32]。低劣数据在全球造成了严重后果,数据错误导致的患者死亡数约占美国全部医疗事故死亡人数的50%^[30]。确保数据可用性是一项十分困难的任务,弱可用数据知识的发掘是其中之一^[22]。数据易用性表现为易见、易学和易用,应贯穿大数据整个流程,可视化、匹配、反馈、人机交互以及数据起源技术可有效地提升易用性。其他问题包括高能耗、不同架构硬件共存和缺乏大数据管理测试基准等。

2.3 大数据发展

大数据导致技术和社会变革,软件开源与数据开放成为潮流。大数据管理需变革与创新,大数据将成为推动IT等产业的发展引擎,催生大数据产业^[24],数据资源服务商将会不断出现^[23,29],大数据发展主要集中在整体发展、学术研究、人文、安全和隐私、应用、系统和处理、大数据对产业影响等7个领域的诸多方面^[24]。但也有人认为,大数据概念的兴起不排除是技术营销的结果,大数据不过是对数据管理挑战认识的新说法^[24]。

3 大数据在医学中的应用

截至2014年8月9日,PubMed数据库文题中含有“big data”的论文共231篇,其中5篇与肿瘤有关^[2,33,34],4篇和流行病学研究有关^[35],仅1篇涉及肿瘤流行病学^[36]。总的而言,PubMed相关论文涉及的学科和内容较多,如免疫^[37]、基因测序^[33,35]、器官移植^[38]、放射生物学^[39]、胃肠道疾病研究^[40]、脑科研究^[41]、高危与高价值患者管理^[42]、卫生统计^[43]和传染病研究^[44],最多的是有关大数据在卫生保健中的应用^[45],电子健康数据正在改变卫生保健、公共卫生行为和临床研究。Curtis等人^[46]描述了共享美国国家大数据网络的可能,重新定位了电子健康记录和管理数据的目的,例举了美国食品和药物管理局前哨项目(the Food and Drug Administration's Mini-Sentinel program)、美国国家临床研究网络、美国国家卫生研究院卫生保健系统研究合作分布研究网络和公共卫生监测(ESPnet)等4个项目,并认为使用这些电子健康数据的挑战包括数据采集、编码、保存、隐私和专利等方面。美国卫生研究所启动了一个大数据知识项目(BD2K),探讨如何更好地提取数据、创造数据应用分析工具和培养人才等^[47]。

中国知网共有大数据相关论文和报道约300篇,仅25篇与医学有关,远少于PubMed的数量,涉及的内容也较少,如探讨大数据在创伤性休克复苏^[48]、个体化医疗^[49]、新PACS系统^[50]、公共卫生和传染病监测^[51]、疾病负担^[52]、中医药评价体系创建^[53]和用户健康信息^[54]中的应用,介绍了如何应用媒体与互联网数据分析预警传染病爆发和发现药物副作用、健康数据服务、虚拟助手如肺癌治疗管理决策^[55]和医学信息管理专业人才培养^[56]等,部分探讨了“可穿戴设备”对健康管理的应用,无有关癌症和肿瘤流行病学研究的文章。

4 肿瘤登记中大数据的应用

肿瘤基础数据具有典型的大数据特征,即大量、结构复杂、变化快和具有价值,且随着医学、科技和肿瘤防治的发展与需要,肿瘤基础数据将会越来越多,越来越复杂,越来越难以用传统肿瘤登记方法收集和整理,必须采用新的如大数据的技术和方法进行应对。肿瘤登记包括资料的收集、整理、存储、编码、统计分析、应用和管理等过程,但瓶颈主要在资料的收集、整理和编码,而资料收集又是其中最关键、最困难和最基本的环节^[14],因而提高肿瘤登记质量最关键的就是如何获取高质量肿瘤资料。经济、医疗卫生和科技尤其是信息技术的发展,为肿瘤资料收集提供了新的方法和手段。广东省中山市肿瘤登记处早就提出利用先进的信息化和智能化技术,创新肿瘤资料收集方法,进行肿瘤资料收集与编码的自动化,进而实现肿瘤登记的全自动化^[57,58]。在全国肿瘤防治研究办公室和全国肿瘤登记中心的指导和协助下,中山市人民医院采用商业智能方法,实现了肿瘤资料收集、编码、统计和不同数据链接的自动化,并正在努力实现资料的自动提取和整合^[8,12]。其研发的肿瘤自动查找软件运行速度快,1 h即可完成1个月的手工资料查找,准确性高,灵敏度和特异性分别达98.91%和99.22%,高于国外同类软件^[9,59]。软件不仅大大减少了肿瘤登记人员的工作量、繁琐,也减少了人力、财力和物力的投入。肿瘤登记自动化迎合了科技、医疗卫生和大数据的发展趋势,不仅可推动肿瘤登记的发展,也可借鉴应用到其他病种如传染病和糖尿病的监测和研究中。肿瘤登记今后的发展方向之一,就是利用大数据的技术和方法,实现肿瘤登记自动化。

大数据技术和方法不仅可推动肿瘤登记自动化,

也可促进人群肿瘤登记和研究的开展。北卡罗莱纳州采用大数据技术开发的综合癌症信息和监测系统(Integrated Cancer Information and Surveillance System, ICISS),将不同数据、方法和系统有机结合起来,可有效地从个人、健康服务商和地区等不同途径,如北卡罗莱纳州癌症登记处、Medicare、Medicaid、美国人口普查、CDC社区卫生指标和私人健康保险计划等,及时、完整和准确地收集、整合和更新肿瘤相关资料。该系统不仅促进了人群肿瘤登记的发展,也为不同领域研究人员开展基于人群的肿瘤研究提供了可能和便利,极大地扩大了肿瘤登记资料的可用和易用性^[36]。所以,肿瘤登记今后发展的第2个方向是采用大数据的技术和方法,及时、完整和准确收集、整合和更新不同来源的肿瘤数据,提高肿瘤资料的质量、可用性和易用性,推动肿瘤登记资料的开放和共享,扩展肿瘤登记资料的应用,如研究环境、经济、文化和个人习惯等对肿瘤流行和防治的影响,更好地为肿瘤防治和流行病学研究服务。

随着国家“3521”工程的不断推进,大量不同的医疗卫生数据被收集、存储在卫生信息平台上,形成了许多包含各种医学信息如肿瘤信息的数据中心,而采用大数据技术和方法建立的地区、省和国家医疗卫生数据中心,大大方便了肿瘤登记开展,肿瘤登记以后可能不再需要与众多的部门和科室打交道,只需在信息平台上获取所需资料即可^[60]。因而,今后肿瘤登记的第3个发展方向是,随着采用大数据技术

和方法建立的各级医疗卫生信息中心的出现,肿瘤登记模式将彻底改变,肿瘤资料将只需从各级信息中心获取即可。

随着肿瘤登记覆盖人群和肿瘤信息的迅速增多、登记水平的不断提高、管理规范化与标准化的迫切需要,国家“3521”信息化工程和其他数据中心的示范效应,建立不同级别如国家和省级的肿瘤数据中心,逐渐成为可能和必要。因而,今后肿瘤登记发展的第4个方向是采用大数据的技术和方法,建立各级肿瘤数据中心。而肿瘤数据中心的建立,将有效地解决肿瘤数据查找、提取、整理、存储、计算、安全、管理和应用等问题,促进肿瘤数据的开放和共享,从而推动肿瘤登记和肿瘤流行病学研究的开展。

随着肿瘤登记方法的改变、肿瘤登记自动化的逐步实现、各级卫生信息平台和数据中心的建立、肿瘤流行病学研究和防治的需要和对肿瘤防治的不断重视,肿瘤登记的第5个发展方向是,大数据技术和方法将推动和促进医院肿瘤登记的开展和人群肿瘤登记的覆盖率,部分经济发达和技术先进的省份将率先开展全省肿瘤登记,而全国肿瘤登记的开展逐渐成为可能。

总而言之,随着经济、科技和社会的发展,尤其是信息技术的发展,大数据发展突飞猛进,而大数据的发展又进一步促进了肿瘤登记的发展,肿瘤登记自动化不仅符合大数据和肿瘤流行病学发展的大趋势,更可借鉴利用到其他疾病的监测与研究中。

参考文献

- 1 Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN 2012 v1.0, cancer incidence and mortality worldwide: IARC Cancer Base No. 11 [OL]. Lyon: International Agency for Research on Cancer, 2013
- 2 Bray F, Znaor A, Cueva P, et al. Planning and Developing Population-based Cancer Registration in Low-and Middle-Income Settings. Lyon: IARC Technical Publication 2014, 43: 1
- 3 Xiao F, Qi L L. Big data processing technics and exploration (in Chinese). Comp Modern, 2013, 217: 75-77 [肖飞, 齐立磊. 大数据处理技术与探索. 计算机与现代化, 2013, 217: 75-77]
- 4 IARC. Global Initiative for Cancer Registry Development. Available at <http://gicr.iarc.fr/>
- 5 Curado M P, Edwards B, Shin H R, et al. Cancer Incidence in Five Continents Volume IX. Lyon: IARC Scientific Publications, 2007: 1
- 6 Sankaranarayanan R, Swaminathan R. Cancer Survival in Africa, Asia, the Carribean and Central America. Lyon: IARC Scientific Publications, 2011, 162
- 7 Timilshina N, Tsomo U, Mklos S, et al. Initiation of Bhutan cancer registry: Experience, progress and future challenge from low-middle income country. In: IACR/NAACCR. 36th International Association of Cancer Registries Annual Conference 2014. Ottawa, Canada, 2014. 47
- 8 Liang Z H, Liu J, Wang Y N, et al. Development and application of automatic cancer data finding software (in Chinese). China Cancer, 2012, 21: 488-490 [梁智恒, 刘静, 王亚娜, 等. 肿瘤资料自动收集软件的研发与应用. 中国肿瘤, 2012, 21: 488-490]
- 9 Hanauer D A, Miela G, Chinnaiyan A M. The registry case finding engine: An automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg, 2007, 205: 690-695

- 10 Tagliabue G, Maghini A, Fabiano I S, et al. Consistency and accuracy of diagnostic cancer codes generated by automated registration: Comparison with manual registration. *Population Health Metrics*, 2006, 4: 10
- 11 Tognazzo I S, Andolfo I A, Bovo I E, et al. Quality control of automatically defined cancer cases by the automated registration system of the Venetian Tumour Registry. Quality control of cancer cases automatically registered. *Eur J Public Health*, 2005, 15: 657–664
- 12 Wei K R, Chen W Q, Liang Z H. Finding and coding cancer data automatically. In: IACR/NAACCR. 36th International Association of Cancer Registries Annual Conference 2014. Ottawa Canada, 2014. 45
- 13 Wei K R, Liang Z H, Liu J, et al. Brief history of Chinese cancer registration (in Chinese). *Chin J Med History*, 2012, 42: 21–25 [魏矿荣, 梁智恒, 刘静, 等. 中国肿瘤登记简史. *中华医史杂志*, 2012, 42: 21–25]
- 14 Wei K R, Chen W Q, Zhang S W, et al. Cancer registration in the Peoples Republic of China. *Asian Pac J Cancer Prev*, 2012, 13: 4209–4214
- 15 Chen W Q, Zhang S W, Zeng H M, et al. Cancer incidence and mortality in China, 2010 (in Chinese). *China Cancer*, 2014, 23: 1–10 [陈万青, 张思维, 曾红梅, 等. 中国 2010 年恶性肿瘤发病与死亡. *中国肿瘤*, 2014, 23: 1–10]
- 16 Forman D, Bray F, Brewster D H, et al. Cancer Incidence in Five Continents, Volume X. Lyon: IARC Scientific Publication [2014-06-18]. <http://ci5.iarc.fr>
- 17 Li Z H, Wang G R, Zhou A Y. Interperating the development and trend of big data from the view of database (in Chinese). *Comp Engin Sci*, 2013, 35: 1–11 [李战怀, 王国仁, 周傲英. 从数据库视角解读大数据的研究进展与趋势. *计算机工程与科学*, 2013, 35: 1–11]
- 18 Li G D. Development of medical informatization in the era of big data (in Chinese). *Silicon Valley*, 2013, 139: 7–8 [李国栋. 大数据时代背景下的医学信息化发展. *硅谷*, 2013, 139: 7–8]
- 19 Meng X F, Ci X. Big data management: Concepts, technics and chanllenge (in Chinese). *Comp Res Dev*, 2013, 50: 146–169 [孟小峰, 慈祥. 大数据管理: 概念、技术与挑战. *计算机研究与发展*, 2013, 50: 146–169]
- 20 Feng D G, Zhang M, Li H. Security and privacy protection of big data (in Chinese). *Chin J Comp*, 2014, 37: 246–258 [冯登国, 张敏, 李昊. 大数据安全与隐私保护. *计算机学报*, 2014, 37: 246–258]
- 21 Wang Y J. Research on the application of information resource in the era of big data (in Chinese). *Sci Inform Dev Economy*, 2013, 23: 124–126 [王玉君. 大数据时代信息资源利用研究. *科技情报开发与经济*, 2013, 23: 124–126]
- 22 Hu X W, Zhang B L, Li D F. Big data research and application (2) (in Chinese). *Standard Sci*, 2013, 10: 18–21 [胡雄伟, 张宝林, 李抵飞. 大数据研究与应用综述(中). *标准科学*, 2013, 10: 18–21]
- 23 Hu X W, Zhang B L, Li D F. Big data research and application (1) (in Chinese). *Standard Sci*, 2013, 9: 29–34 [胡雄伟, 张宝林, 李抵飞. 大数据研究与应用综述(上). *标准科学*, 2013, 9: 29–34]
- 24 Hu X W, Zhang B L, Li D F. Big data research and application (3) (in Chinese). *Standard Sci*, 2013, 11: 29–33 [胡雄伟, 张宝林, 李抵飞. 大数据研究与应用综述(下). *标准科学*, 2013, 11: 29–33]
- 25 Liang F L. Research on the ways of managing data in the era of big data (in Chinese). *Comp Dev Appl*, 2013, 26: 34–36 [梁凤兰. 大数据时代数据管理方式的研究. *电脑开发与应用*, 2013, 26: 34–36]
- 26 Reportbuyer. The Big Data Market: 2014-2020 Opportunities, Challenges, Strategies, Industry Verticals and Forecasts. [2014-06-18]. <https://www.reportbuyer.com/product/2164339>
- 27 Liu L, Yan Y J. Research on the strategy of Australia public service big data (in Chinese). *Library Sci Res*, 2014, 5: 47–52 [刘兰, 闫永君. 澳大利亚公共服务大数据战略研究. *图书馆学研究*, 2014, 5: 47–52]
- 28 Li K. Big data: From conceptualization to Valuation (in Chinese). *China Financ Comp*, 2014, 3: 19–20 [李克. 大数据: 从概念化走向价值化. *中国金融电脑*, 2014, 3: 19–20]
- 29 Chen X Y. Business value of big data (in Chinese). *Business Manag*, 2013, 3: 108–110 [陈宪宇. 大数据的商业价值. *企业管理*, 2013, 3: 108–110]
- 30 Dong J H. Trend following-how HIT in USA meet the big data era (in Chinese). *Med Inform J*, 2013, 34: 2–6 [董建华. 顺势而行—美国 HIT 如何迎接大数据时代的来临. *医学信息杂志*, 2013, 34: 2–6]
- 31 Bai R J, Leng F H. Research on the data integration in the era of “big data” (in Chinese). *Pract Res*, 2014, 37: 94–98 [白如江, 冷伏海. “大数据”时代科学数据整合研究. *实践研究*, 2014, 37: 94–98]
- 32 Li J Z, Liu X M. An important aspect of big data: data availability (in Chinese). *Comp Res Dev*, 2013, 50: 1147–1162 [李建中, 刘显敏. 大数据的一个重要方面: 数据可用性. *计算机研究与发展*, 2013, 50: 1147–1162]
- 33 Phillips K A, Trosman J R, Kelley R K, et al. Genomic sequencing: Assessing the health care system, policy, and big-data implications. *Health Aff (Millwood)*, 2014, 33: 1246–1253
- 34 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
- 35 Kao R R, Haydon D T, Lycett S J, et al. Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends Microbiol*, 2014, 22: 282–291

- 36 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
- 37 Zhang G L, Sun J, Chitkushev L, et al. Big data analytics in immunology: A knowledge-based approach. *Biomed Res Int*, 2014, 2014: 1–9
- 38 Massie A B, Kuricka L M, Segev D L. Big Data in organ transplantation: Registries and administrative claims. *Am J Transplant*, 2014, 14: 1723–1730
- 39 Rosenstein B S, West C M, Bentzen S M, et al. Radiogenomics: Radiobiology enters the era of big data and team science. *Int J Radiat Oncol Biol Phys*, 2014, 89: 709–713
- 40 Choi J, Park H, Lee C H. Analysis of a blog for gastrointestinal disease the viewpoint of the big data: A single institutional study. *Kor J Gastroenterol*, 2014, 63: 361–365
- 41 Choudhury S, Fishman J R, McGowan M L, et al. Big data, open science and the brain: Lessons learned from genomics. *HumNeurosci*, 2014, 8: 239
- 42 Bates D W, Saria S, Ohno-Machado L, et al. Big data in healthcare: Using analytics to identify and manage high-risk and high-cost patients. *HealthAff (Millwood)*, 2014, 33: 1123–1131
- 43 Yoo C, Ramirez L, Liuzzi J. Big data analysis using modern statistical and machine learning methods in medicine. *Int Neurourol J*, 2014, 18: 50–57
- 44 Dutruel C, Thole J, Geels M, et al. TRANSVAC workshop on standardisation and harmonisation of analytical platforms for HIV, TB and malariavaccines: How can big data help? *Vaccine*, 2014, 32: 4365–4368
- 45 Bellazzi R. Big data and biomedical informatics: A challenging opportunity. *Yearb Med Inform*, 2014, 9: 8–13
- 46 Curtis L H, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data a network. *HealthAff (Millwood)*, 2014, 33: 1178–1186
- 47 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 InformAssoc*, 2014, 21: 957–958
- 48 Zeng J, Yang H, Jiang H. Challenge and opportunity of research on the traumatic shock resuscitation: integration of big data analysis, computer science, systematic biology and trauma science (in Chinese). *Traum Surg J*, 2013, 15: 186–189 [曾俊, 杨浩, 江华. 创伤性休克复苏研究的挑战与机遇: 大数据分析、计算机科学、系统生物学与创伤科学的融合. *创伤外科杂志*, 2013, 15: 186–189]
- 49 Li J, Gu J. Individualized medicine and the opportunity and challenges in the age of big data (in Chinese). *Med Philosophy*, 2014, 35: 5–11 [李静, 顾江. 个体化医疗和大数据时代的机遇和挑战. *医学与哲学*, 2014, 35: 5–11]
- 50 Xiang H Q. New PACS system to response to the big hospital imaging transmission—a recoding of EBM Hybrid 64 PACS system test (in Chinese). *Hospital Inform*, 2014, 3: 64 [相海泉. 新 PACS 系统应对医院大数据量影像传输——记 EBM Hybrid 64 位 PACS 系统测试. *医院信息化*, 2014, 3: 64]
- 51 Xiang H Q. Big data application in public hygiene-interview with the director of information center of Chinese Disease Control and Prevention (in Chinese). *Inform China*, 2013, 5: 44–45 [相海泉. 公共卫生的大数据应用——专访中国疾病预防控制中心信息中心主任马家奇. *中国信息界*, 2013, 5: 44–45]
- 52 Yu S C, Xiao G X. Global disease burden—An example of big data analytical application (in Chinese). *Med Inform J*, 2013, 34: 12–16 [于石成, 肖革新. 全球疾病负担研究——大数据分析应用实例. *医学信息学杂志*, 2013, 34: 12–16]
- 53 World Chinese Traditional Medicine and Drug. Hugging big data era, creating evaluating clinical system of Chinese traditional medicine and drugs-interview with executive vice president of China Academy of Chinese Medical Science (in Chinese). *World Chin Tradit Med Drug*, 2013, 8: Front page 1–2 [世界中医药. 拥抱大数据时代, 创建中医药临床评价体系——访中国中医科学院刘保延常务副院长. *世界中医药*, 2013, 8: 前插 1–前插 2]
- 54 Peng Y, Yan L, Zhu H. The value of user health informatics in the age of big data (in Chinese). *Med Inform J*, 2014, 35: 2–6 [彭琰, 严莉, 朱红. 大数据时代用户健康信息学的价值. *医学信息学杂志*, 2014, 35: 2–6]
- 55 Zhang C, Li X. When big data encounter medicine (Series 2) (in Chinese). *Inform China*, 2013, 8: 58–63 [张弛, 李翔. 当大数据与医疗“邂逅”(连载二). *中国信息界*, 2013, 8: 58–63]
- 56 Wang X D, Wu Y Q. Training of medical information management professionals in the age of big data (in Chinese). *Med Inform J*, 2014, 35: 83–85 [王晓东, 吴雅琴. 大数据时代医学信息管理专业人才培养. *医学信息学杂志*, 2014, 35: 83–85]
- 57 Wei K R, Liang Z H, Ren X Q. Innovation of cancer data collecting ways (in Chinese). *China Cancer*, 2009, 18: 528–530 [魏矿荣, 梁智恒, 任小青. 肿瘤资料搜集方法的创新. *中国肿瘤*, 2009, 18: 528–530]
- 58 Liang Z H, Liu J, Wei K R. Some opinions about cancer registration (in Chinese). *China Cancer*, 2010, 12: 779–781 [梁智恒, 刘静, 魏矿荣. 对肿瘤登记工作的一些看法. *中国肿瘤*, 2010, 12: 779–781]
- 59 Contieroa P, Tittarellia A, Maghini A, et al. Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system. *J Biomed Informatics*, 2008, 41: 24–32

- 60 Cai J H, Zhang T, Zong W H. Challenge of medical big data and reflection (in Chinese). *Chin J Health Inform Manag*, 2013, 10: 292–295 [蔡佳慧, 张涛, 宗文红. 医疗大数据面临的挑战及思考. 中国卫生信息管理杂志, 2013, 10: 292–295]

Influence of “big data” on the development of cancer registration methods

WEI KuangRong¹, LIU ShenChao², WEI DongLin³, LIANG ZhiHeng¹ & CHEN WanQing⁴

¹ *Cancer Institute, Zhongshan People's Hospital, Zhongshan 528403, China;*

² *Department of Technology, Dimensional Insight, Massachusetts 01803, USA;*

³ *Department of Computer Science, Dartmouth College, New Hampshire 03755, USA;*

⁴ *National Office for Cancer Prevention and Control & National Central Cancer Registry, Beijing 100021, China*

The development of cancer registration methods and use of “big data” in China, and the influence of each upon the other, were reviewed. It was found that there have been recent, rapid development in big data and cancer registration, with each having influenced development of the other. Under the influence of big data, cancer registration methods are likely to develop in the following ways: (1) Automated cancer registration; (2) timely collection and collation of cancer data from various sources; (3) overhaul of cancer registration methods with the emergence of health information centers at different levels; (4) construction of cancer data centers at different levels using big data techniques; (5) gradual uptake of hospital-based cancer registration and rapid growth of population-based registration, leading to coverage of the entire population in some more developed provinces. These developments will improve cancer data quality, availability and usability, and will facilitate data sharing, thus extending the application of cancer registration data in areas including surveillance and research into other diseases.

big data, cancer registration, epidemiology

doi: 10.1360/N972014-00882