

医疗大数据建设现状及其应用发展对策研究

刘文韬¹, 王仁佐²

(1. 上海交通大学医学院附属瑞金医院, 上海 200025; 2. 上海森亿医疗科技有限公司, 上海 201203)

摘要: 医疗大数据产业是全球竞相建立和拓展的产业, 它的发展为全民健康素质、疾病防治、药物研发等相关健康产业的发展提升, 提供了非常重要的基础和支持, 是大数据产业中非常重要的组成部分。中国医疗大数据具有资源丰富、涉及面广等优势, 但医疗地区发展不平衡、医疗数据资源无法共享、数据资料质量不统一等缺陷, 极大阻碍了中国医疗大数据产业的发展。所以建立医疗大数据人才库、尽快共享和整合医院医疗数据库、将医疗数据结构化并建立质控标准、指导医疗大数据的安全应用等是当前中国医疗大数据产业健康快速发展的当务之急。

关键词: 医疗大数据; 人工智能; 预测医学; 大数据产业

中图分类号: R19 **文献标识码:** A **文章编号:** 1671-2404(2018)86-0080-04

《自然》杂志在 2008 年 9 月推出了名为“大数据”的封面专栏后, 大数据的概念广泛引起各个专业领域的高度重视。而在维克托·迈尔·舍恩伯于其著作《大数据时代》中首次明确定义大数据的 4V 特点: Volume(数据量大)、Velocity(输入和处理速度快)、Variety(数据多样性)、Value(价值密度低), 并且随着计算机技术发展出现云计算后, 大数据逐步展现其真正价值。

当前大数据和云计算为社会管理、医疗、科研、教育都带来了前所未有的发展契机。2012 年, 美国奥巴马政府在白宫网站上发布了《大数据研究和发​​展倡议》, 并投入 2 亿美元以上, 已形成了包括联邦政府 12 个部门和机构的多项研究计划。中国国务院相继印发了《关于印发促进大数据发展行动纲要的通知》和《关于促进和规范健康医疗大数据应用发展的指导意见》。近期, 第九届全球健康促进大会发布《上海宣言》《健康城市上海共识》, 为全球推进健康城市建设、实现可持续发展目标提供了方向。随着计算机技术的发展, 大数据产业在全球方兴未艾, 中国多地筹集的科创中心, 进行的大数据研发也是新兴企业及创业的重点内容。比如, 在上海科创中心启动并建设中, 大数据研发服务是科创中心的重点之一。其中疾病和健康大数据不仅对医学科研和临床治疗有巨大价值, 而且这些数据的研发对于

健康策略和政策的制定也有重大意义, 所以开发健康及病例资料大数据具有巨大的科研价值和应用价值。

1 医疗大数据数据库建设的紧迫性及其应用价值举例

1.1 医疗大数据与临床研究及药品研发

随着大数据开发, 计算机技术应用及人工智能的发展, 使得医疗大数据为基础的健康研究和发​​展成为可能, 全国各地都在抢占该产业的先机。2016 年 12 月 25 日, 美国国会在其官方网站上公布了《21 世纪治愈法案(21st Century Cures Act)》最终版本, 该法案从提出到一次次修改, 再到最终版推出, 历经两年多时间, 终于让这一推动美国未来 10 年或更长时间内生物医学创新研发、疾病治疗及大健康领域发展的法案落地生根。而法案中关于利用“真实世界证据”(Real World Evidence, 以下简称 RWE) 取代传统临床试验进行扩大适应证的批准, 引起了极大关注。“真实世界数据”的最重要的核心内涵就是基于人群大数据的研究, 从而推动药品研发和应用以及建立治疗方案疗效客观评价的新方法。通过人群健康数据及医疗大数据对药品进行评估只是医疗大数据最简单的应用。FDA 于 2017 年 9 月进一步提出《真实世界证据支持医疗监管决策指南》, 明确指出: “如 RWD 所包含的临床数据中产生的高质量 RWE, 能够提供或支持对产品全生命周期进行管理所需要的信息分析, FDA 将考虑采纳 RWE 来支持医疗器械的审批决策”, 并进一步对大数据的应

收稿日期: 2017-11-01

作者简介: 刘文韬, 主治医师, 博士, 主要从事肿瘤大数据及肿瘤诊治研究; 王仁佐, 高级工程师, 主要从事人工智能医学自然语言处理及大数据治理和应用工作。E-mail: wt_mygod@163.com

用场景及要求进行规范。

1.2 医疗大数据与人工智能

结合人工智能的医疗大数据可以提供更广泛的应用方法。一个医疗大数据结合人工智能的简单应用就是 IBM Watson for Oncology,其通过对大量临床研究数据的采集和分析,自动提供肿瘤患者的治疗方案。该机器人已经引进中国近二十余家大型医院,并开始对患者进行辅助诊治。这项技术只是医疗大数据与人工智能的最简单结合。但是在比较复杂的系统中,大数据挖掘和人工智能可能会受挫。近期,据福布斯的报道指出,IBM 与该世界顶尖癌症研究机构的合作关系正趋于破裂。此前安德森癌症中心证实:此项目从去年开始就已经暂停。安德森癌症中心也正在积极寻求其他合作方的竞价,未来这些合作方有可能取代 IBM。来自德克萨斯大学审计机构的一份报告指出,安德森癌症中心已经花费了 6 200 万美金用于此项目,但尚未实现目标。审计记录显示项目重点更换了数次,第一次重点研究白血病,然后又是肺癌,最后毫无进展。

1.3 医疗大数据与预测医学

预测医学是二十世纪八十年代新兴的一门综合性学科,是以预知先觉人体身心健康与病症为主,为人类健康服务的医学科学体系。疾病预测采取病因预测、超前期预测、症状前预测和临床预测等四级预测措施。测治于未萌,防病于未然,泯灭在发作之前。上世纪兴起的基因组、蛋白质组学等通过高通量技术预测疾病发生发展获得了巨大的关注,对其研究推动了对大数据数理模型的技术发展。越来越多的研究发现疾病不仅取决于基因及其表达还与表观修饰蛋白功能等多方面协同作用,仅对基因组等单一组学进行微观研究较难获得突破。医疗大数据的发展使得微观基因组基础大数据和宏观临床大数据结合,联合信息统计分析人工智能技术,可更加准确地预测疾病的发生发展,从而推动预测医学的进一步提高。2015 年奥巴马宣布一项名为“精准医学”的计划,耗资 2.15 亿美元打造 100 多万名美国志愿者的基因信息库进行分析,更好地了解疾病形成机理,进而为开发相应药物、实现“精准施药”铺平道路。

2 中国医疗大数据建设的优势及劣势

中国医疗资源丰富但信息化起步较晚,患者疾

病及健康数据库的建设十分滞后,医疗大数据的研发应用产品十分缺乏。

2.1 中国医疗数据库建设

中国人口众多,医疗卫生资源分布不均,大医院较多集中在中大型城市,这些医院病种丰富、患者众多,除当地患者外还吸引从全国各地慕名前来求医的患者。由于近些年国家交通等基础设施建设的投入,重大疾病患者多可方便自由选择大城市中大型医院,而基本上中大型医院也对每位患者就诊及住院治疗时均建立了完整的检查及病例资料。许多医学中心早在 10 年前就建立了患者电子病历资料数据。比如作者所在的上海交通大学医学院附属瑞金医院胃肠外科,近十年已累计胃癌外科治疗患者近万例。胃癌单病种累计医疗数据就是欧美等多国联合中心积累几十年的病例。

虽然中国由于人口众多,患者数据量大,形成医疗大数据的天然基础。但是劣势也非常明显。由于缺乏统一规划,原始病例资料数据庞大,分散在各级医院各个不同的业务系统中,虽然已经形成电子病史系统,但各系统间数据整合不易。加上数据所有权尚存争议,无相关的政策法律定义及支持,区域、地域间数据多无共享,无法形成可利用的大数据共享资源供研究分析,更加无法建设产品应用。

2.2 中国医疗数据库资料质量及结构化状况

中国医疗数据由于早期计算机信息技术的发展,大部分医院都已通过信息系统进行管理和应用。但是医疗数据资料较为复杂,包括门诊诊疗数据,病史数据,实验室检查结果,辅助检查结果,手术及治疗数据等等,内容众多,形式多样,结构复杂。虽然医院内是电子化,但早期病例源文件保存多为 PDF 等图片格式保存,未形成可分析利用的点位值数据源。即使有部分结果如化验报告已经形成可以利用的点位数据值,但多不全面,无法进行深度发掘和整合。医疗数据碎片化及无结构化是医疗电子数据的特点也是医疗大数据应用的难点。未结构化的数据实际进行挖掘的利用率低,有待引入自然语言处理技术完成结构化及进一步设计新的应用挖掘方法解决。

2.3 中国医疗大数据资源共享及利用情况

中国各大医院医疗数据结果无法共享,是目前医疗数据库面临的重大问题。每个医院有各自的病史病例系统,彼此不统一,各成系统,无法直接共享。

由于历史或管理原因,甚至有些医院内部不同部门之间的医疗信息都无法实行共享。且因包含隐私信息,医院及上级管理部门为避免麻烦,也不愿意进行共享。由此也进一步限制了第三方企业或机构进行数据采集,限制了医疗数据研发企业的创业和创新发展,导致这些信息及病史数据无法变为大样本的大数据网络供研发。

2.4 中国医疗大数据人才情况

中国由于后发优势,计算机及信息化水平并不低,而且中国医疗大数据应用的主体医护人员素质和能力整体水平也较高。由于信息化网络基础设施建设发展,人们对信息化及数据的应用并不陌生,且已初步形成大数据研究的专业化人员队伍。但是医疗大数据建设及其研发有特殊性。首先医疗数据专业性强,其采集、整合、解读和应用都需要专业化卫生人员甚至医护人员的参与。然而中国了解医疗大数据及参与医疗大数据建设的卫生人员极其稀缺,此类人才不光需要有医学背景,而且需要具有数理统计和信息分析等知识。中国已建设的大数据人才多是数理统计方面的专家,但由于医学知识的缺乏,就无法真正深度应用已整合的医疗数据。所以医数医理交叉人才的短缺及医护人员大数据相关知识的缺乏是目前阻挡中国医疗大数据发展的重大问题。医疗大数据相关研究基金和课题的缺乏也是目前卫生人员对医疗大数据不了解和不深入研究的原因之一。

3 中国医疗大数据建设及应用对策

由于中国医疗大数据建设存在结构化水平低、共享性差,专业化管理研究人员少等问题,需出台多项政策及相关资金扶持产业快速健康发展。

3.1 结构化水平和数据质量是医疗大数据的生命

随着医学进入信息和大数据时代,由于时间地域及硬件条件的不平衡,可能生成海量的医疗病例错误信息。有误的病历电子化后,将带来更多更大的扭曲删改和误导性的信息。此外,数据的不完整性也是制约医疗大数据发展的因素,可能对其应用产生错误或偏倚。所以,尽快统一各级医院病例数据模块,建立标准化的输出结构或格式,使得数据的采集标准化、规范化,让数据能够分析使用是当务之急。不同类型的医疗数据资料,采集传输时达到统一标准。收集规范化资料建立患者医疗大数据库,

可共享使用。此外更重要的是建立针对科研或医疗大数据应用的医疗大数据质量评估方法和体系。建立质量评估评分体系,一方面可以促进医疗大数据质量的提高,另一方面根据分数和分级可区分不同应用层面和应用领域,促进其科学应用。同时创建中国科学的医疗大数据标准后,可早日规范和指导全球医疗大数据平台建设合作和应用,不受欧美等国家垄断及牵制。

3.2 隐私保护及脱敏对医疗大数据库安全保护

中国医疗大数据目前尚未形成统一共享的体系,最主要的问题是出于医院及相关部门对医疗大数据隐私内容及保护数据库安全考虑。尽快制定相关政策法规,消除医疗数据关于隐私泄露方面限制,进行脱敏处理,使得绝大部分非隐私的个人健康数据得到充分利用,推动健康产业的数据研发。在敏感和隐私信息脱敏的技术层面处理并不复杂,但是相关法律法规和政策的制定是当务之急,以及医疗大数据需根据不同应用需求建立不同权限限制。研究出台医疗大数据共享方案是使得埋于地下的数据金矿得以应用的关键。例如可首先通过各个医院整合其内部医疗数据,再通过各地卫计委统一领导下收集整理上传各个医院的医疗大数据进行脱敏处理后,进行相关研究。也可以先由各医院进行脱敏处理后,在卫计委领导下第三方公司进行收集整理工作,最后形成医疗大数据库。此外,经费可适当向医疗大数据建设和研发产业倾斜,促进其健康发展。

3.3 建立医疗大数据研发企业及人才队伍建设

加快医疗大数据建设可以创建许多医疗大数据研发企业中心,可以为健康事业提供产品,促进全面健康的发展。医疗大数据研发对基础设施建设无需过大投入,初创企业门槛较低,而且医疗大数据研发项目较多,研究方向较广,从采集、处理到应用,涉及多方面内容,均可成立不同小微企业专业化经营,生产不同方向的大数据应用产品。但是医疗大数据相关企业对人才需求较为迫切,尤其是数学、统计、数据处理等方面人才,中国起步较晚,但并非落后。由于中国在企业大数据等领域已积累了一定经验,这推动了这些年来在数据处理方面相关的人才储备。医疗大数据其具有自然的特点,就是有一定专业性,所以国家需重视相关的医理交叉人才的培养。

3.4 指导和普及医疗大数据库的应用

医疗大数据应用领域较为广泛,其对医疗政策

制定、药品研发、科学研究、健康产业发展等方面均有较大价值。然而由于其应用具有一定的专业性,所以需要制定相关政策使临床医师、公共卫生人员可共享脱敏后大数据资料,进行相关研究,并推动卫生工作者与统计师、IT 工程师及社会研究和管理者共同进行相关健康数据产品研发。建立和投入医疗大数据研究相关基金,支持该产业产品应用研发,并且指导相关人员进行产品应用,可推动该产业发展。

4 总结及展望

医疗大数据覆盖面广,数据量大,数据结构复杂,来源广泛,应用面广,专业性强,所以医疗大数据工程需要顶层设计,加快相关法规和规范的出台。医疗大数据安全是其应用的保证,结构化和质量可靠的大数据库建设需要建立大数据质量评估体系推动。目前必须认清它存在的缺陷和困难,找到合适的采集和应用方法,而不是一味吹捧,才能真正应用医疗大数据这个数据财富,做到弯道超车,创造医疗大数据全球产业高峰。

参考文献

- [1] Buxton B, Hayward V, Pearson I, et al. Big data: the next Google[J]. *Nature*, 2008, 455(7209):8-9.
- [2] Viktor Mayer-Schö nberger. Big data: a revolution that will transform our lives [J]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, 2015, 58(8): 788-793.

- [3] Tom K. Big Data Is a Big Deal. [EB/OL]. [2014-03-21]. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.Pdf.
- [4] Food and Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices: draft guidance for industry and Food and Drug Administration staff [EB/OL]. (2016-7-27) <http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/>.
- [5] 划时代的意义: IBM Watson 正式落地中国 21 家医院 [EB/OL]. <http://www.itmsc.cn/archives/view-115982-1.html>.
- [6] Special Review of Procurement Procedures Related to the M. D. Anderson Cancer Center Oncology Expert Advisor Project [EB/OL]. <https://www.utsystem.edu/documents/docs/financial/utmdacc-special-review-procurement-procedures-related-oncology-expert>.
- [7] 马慰国, 胡广群. 遗传病的预测医学特点 [J]. *中国优生与遗传杂志*, 2000, 8(4): 119-121.
- [8] The Precision Medicine Initiative [EB/OL]. <https://obamawhitehouse.archives.gov/node/333101>.
- [9] 林琳, 白波, 王韬. 临床数据中心的构建与应用 [J]. *中国数字医学*, 2016, 11(8).
- [10] 洪弘, 李玲娟. 医疗数据挖掘的特点、过程及方法 [J]. *价值工程*, 2011, 30(32).
- [11] Weiskopf NG1, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use [J]. *J Biomed Inform*, 2013, 46(5): 830-6.

Research on Big Data Construction Status and Its Application and Development Strategy

LIU Wentao¹, WANG Renzuo²

(1. Ruijin Hospital Shanghai Jiaotong University School of Medicine, Shanghai Province 200025, China;

2. Shanghai Senyi Medical Technology Ltd Company, Shanghai Province 201203, China)

Abstract: Medical big data industry is one that is fiercely established and expanded globally, whose development offers essential basis and support for the health of all people, disease prevention and treatment, and medicine research, and plays a key part in big data industry. Chinese big data is reputed for its richness and wide range. However, there are many barriers lying ahead which impede China's big data industry, such as the imbalanced medical development between different regions, the inability of medical data sharing and so on. Therefore, it is imperative to build a talent bank for medical big data, share and integrate medical data bank, structuralize the medical data and set quality criteria and guide its safe application, which may help foster the healthy and quick development of China's big data industry.

Key words: Medical big data; artificial intelligence; prediction medicine; big data industry