

·综述·

健康大数据在慢性病预防控制中的应用

刘杨^{1,2} 李辉³ 曾新颖² 董文兰¹ 刘世炜^{1,2}

¹中国疾病预防控制中心慢性非传染性疾病预防控制中心,北京100050; ²中国疾病预防控制中心控烟办公室,北京100050; ³宁波市疾病预防控制中心315010

通信作者:刘世炜, Email: shiwei_liu@aliyun.com

【摘要】 随着信息化的不断发展,大数据(big data)正在越来越多地被运用于慢性病预防控制领域,对居民健康产生重大且深远的影响。本文简要地介绍了大数据和健康大数据(healthy big data)的定义、特征和分类;重点阐述了健康大数据的分析方法及其在慢性病三级预防中的应用;健康大数据应用所面临的技术难题、数据管理和共享、数据质量、伦理和隐私等诸多方面的挑战。为健康大数据在慢性病预防控制方面提供更多研究思路。

【关键词】 健康大数据; 慢性病; 三级预防

基金项目:国家重点研发计划(2017YFC1310902);国家自然科学基金面上项目(81872721)

DOI: 10.3760/cma.j.cn112338-20191119-00815

Application of healthy big data in prevention and control of chronic diseases

Liu Yang^{1,2}, Li Hui³, Zeng Xinying², Dong Wenlan¹, Liu Shiwei^{1,2}

¹National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 100050, China; ²Tobacco Control office,Chinese Center for Disease Control and Prevention, Beijing 100050, China; ³Ningbo Prefectural Center for Disease Control and Prevention, Ningbo 315010, China

Corresponding author: Liu Shiwei Email: shiwei_liu@aliyun.com

【Abstract】 With the continuous development of informatization, big data has been increasingly used in the prevention and control of chronic diseases, which has a significant and considerable influence on public health. This paper briefly introduces the definition, characteristics and classification of big data and healthy big data, focusing on the analysis methods and their applications in tertiary prevention, as well as the challenges in technology, data management, sharing and quality, ethics and privacy, with the aim of providing more research approaches for healthy big data application in chronic disease prevention and control.

【Key words】 Healthy big data; Chronic disease; Tertiary prevention

Fund programs: National Key R&D Program of China (2017YFC1310902); National Natural Science Foundation of China (81872721)

DOI:10.3760/cma.j.cn112338-20191119-00815

2017年全球疾病负担报告指出:慢性病的死亡占全球死亡的73.41%(72.53%~74.06%),伤残调整寿命年(DALYs)占全球的62.02%(60.33%~63.80%);中国慢性病造成的死亡占总死亡的89.48%(89.26%~89.85%),DALYs占比为82.63%(81.99%~83.32%);且根据1990—2017年的变化趋势,可以发现慢性病造成的死亡和DALYs一直呈上升趋势^[1]。虽然随着医疗技术的不断发展,越来越多的治疗手段和药品被应用到慢性病患者治疗过程中,但是中国慢性病的发病率和患病率仍呈现上升趋势,而且还有数量庞大的慢性病潜在患者正处于高危阶段。因此,如何避免高危人群发展为慢性病患者,如何控制慢性病患者的病情进展,如何减少慢性病带来的伤残,是摆在公共卫生人员预防控制慢性病面前的必须要克服的难题。

2016年6月,国务院印发《关于促进和规范健康医疗大

数据应用发展的指导意见》(国办发[2016]47号),指出要夯实健康医疗大数据应用基础,推进公共卫生大数据应用,认为健康医疗大数据应用发展将带来健康医疗模式的深刻变化,有利于激发深化医药卫生体制改革的动力与活力,提升健康医疗服务效率和质量^[2],《“健康中国2030”规划纲要》也明确提出要“推进健康医疗大数据应用”^[3]。健康大数据的快速发展为慢性病的预防控制打开了新的开阔前景。近些年来,健康大数据在慢性病领域的相关研究在国内外正在广泛开展,比如加拿大将初级保健电子病历与人口普查数据相结合,用于研究慢性病的决定因素^[4];英国利用全国大数据中的电子健康记录来帮助改善心血管疾病的预后^[5];浙江省宁波市鄞州区利用区域健康数据研究中国大陆高血压患者使用他汀类药物与新发糖尿病之间的关系^[6]。尽管当前健康大数据的研究十分火热,但仍然不能忽视其中所存在的问

题和困难,诸如信息“孤岛”现象、数据质量问题、伦理隐私问题等^[7-8],这些问题的存在也限制着健康大数据的进一步高质量发展。

本文对健康大数据在慢性病三级预防领域的应用进行梳理,试图找出健康大数据在慢性病防控中的应用发展方向和前景,并对存在的困难进行讨论。

一、大数据的定义及特征

维基百科将大数据定义为:一系列数据集的集合,而这个集合是非常庞大且复杂的,现有的数据库管理工具或传统的数据处理应用程序难以处理。大数据还被定义为以高容量和多样性生成并且以高速累积的数字数据(digital data),导致数据集相对于传统数据处理系统而言太过庞大^[9]。总的来说,大数据是极为庞大且复杂的数据集合,传统的数据分析程序很难对其进行处理。

大数据的特征方面,有研究者曾提出“4V”特征:海量(volume)、多样(variety)、快速(velocity)、价值(value);还有研究者提出“5V”特征:海量(volume)、快速(velocity)、多样(variety)、真实(veracity)、价值(value);还有其他“5V”特征:海量(volume)、多样(variety)、快速(velocity)、可变(variability)、真实(veracity)^[8,10-11]。不论是哪种特征,都阐明了大数据具有数量大、种类多、速度快、数据真实和研究价值高等特点。

二、健康大数据的定义及分类

健康大数据又称为医疗健康大数据,作为大数据系统中的重要组成部分,是医疗卫生领域中的宝贵资源,是以医疗卫生领域为基点进行论述的,包括与维持机体健康相关的生活行为方式、遗传、社会环境因素和医疗过程中的相关数据信息等^[12-13],且涵盖范围广泛,贯穿人的整个生命周期,既包括个人健康,又涉及医药服务、疾病防控、健康保障和食品安全、养生保健等多方面数据的汇聚^[3]。除了具有大数据的“5V”特点外,还包括时序性、隐私性、不完整性等医疗领域的独有特征^[14]。

目前学术界还缺乏一个统一且公认的健康大数据分类系统,根据数据收集渠道的差异,研究人员使用的大数据多数可以分为:①生物学测量的数据,如基因组学或代谢组学数据集;②背景基线测量数据;③行政收集的医疗记录数据,如医疗保险记录的数据;④通过全球定位系统(GPS)设备、运动手环、计步器等以极其频繁的间隔自动进行的测量数据;⑤网络世界中的数据,例如搜索历史记录,社交媒体帖子或手机记录;⑥其他数据,如气象、舆情、疫情、农作物和食品安全等数据^[15]。有研究者提出不同类型的数据对研究和实践的影响并不相同,比如,①、②类数据每一次测量的质量都会受到每次测量时的环境影响,研究分析时需要平衡混杂因素的影响;③、④类数据中存在大量的无关联的或质量很低的数据,分析时需要将这些变量去除;⑤、⑥类数据的直接健康指标很少,很难直接衡量健康水平^[15-16]。

三、健康大数据的分析方法

大数据往往需要经过相关处理后才能凸显其潜在价值。

目前在健康大数据分析上,数据挖掘与传统统计分析两大类方法共存,二者相互补充。此外,人工智能如自然语言处理、模式识别、机器学习等新方法也逐渐用于大数据分析中^[10]。

1. 数据挖掘:数据挖掘是指从大量、不完全、有噪声、模糊、随机的数据中提取蕴含在其中的、事先不得知但又潜在有用的信息和知识的过程^[17]。数据挖掘包括预测性和描述性算法,前者产生用于预测和分类的模型,后者用于发现数据中的关联、聚集和亚组等关系^[18]。公共卫生领域常用的数据挖掘方法有关联分析、分类与预测、聚类分析、决策树、可视化分析、联机分析处理等^[19]。比如,李照东和吴建林^[20]将关联规则用于电子病历数据挖掘,把武汉市某三甲医院近年的数万条记录作为数据源,通过Python数据分析工具,利用关联规则算法对相关的病症和特征进行深入的分析研究,在辅助医疗领域进行尝试和探索。章涛等^[21]收集2013—2014年浙江省11家流感监测哨点医院的检测数据以及气象和空气污染物等资料,通过时滞相关性分析筛选纳入模型的变量,采用2013年第14周至2014年第44周的数据建立Elman神经网络预测模型,采用2014年第45—52周的数据检验模型的预测效能,模型预测效果较好,适用于浙江省流感疫情短期预测。王董磊等^[22]应用Vensim PLE软件构建犬群传播的动力学模型,Tree Age pro 2011软件构建犬咬伤人群免疫策略的决策树模型,通过文献检索和现场调查获得相关概率以及成本数据,研究发现提高犬群疫苗接种覆盖率而减少不必要的人用疫苗,可有利于进一步防控狂犬病且节约社会成本。数据挖掘的工具和方法有很多,需要根据数据特点和研究目的选择合适的分析方法,也常会将几种方法结合使用,弥补各自不足或比较选出最佳方法。

2. 计算技术:为了满足数据的大规模处理需求,一般还需要应用非关系数据库、云计算、云存储等技术对健康大数据进行挖掘、处理和利用,在很多情况下是多种技术被联合使用,如人工智能与并行计算平台的联合使用,或与一些大数据挖掘技术联合使用^[10]。云计算是通过使计算能力分布在大量的分布式计算机集群上,计算能力甚至可以带到万亿次的级别;云存储是在云计算概念的基础上扩展出来的一个新的概念,是指使用集群应用、网格技术或分布式文件系统等功能,将大量不同类型的存储设备通过软件集成起来协同工作,对外提供数据存储服务和业务访问功能的一个系统^[23]。数据仓库,是为决策制定过程提供所有类型数据支持的集合,出于分析性报告和决策支持目的而创建的。随着大数据的不断发展,开发者在传统的数据源基础上,引入了来自于传感器、地理信息、社交网络等多方面的非关系型数据,通过Hadoop软件进行数据处理,运用数据虚拟化技术可以将不同的数据源进行整合,再利用压缩技术来管理更大规模的数据,从而进一步提供数据分析^[24]。

3. 机器学习:传统的方法和机器学习之间的关键区别在于,在机器学习中,模型是从各式案例中学习得来的,而不是用规则进行编程运算。例如,载玻片原本是由病理学家来读取的,但现在可以通过机器学习将载玻片的特征(如载玻片

的像素)转换为一些标签(如指示载玻片包含指示癌症变化的证据的信息)计算机可以使用算法从各种案例中学习,创建一个模型,这个模型可以概括信息,从而可以正确地执行任务(如分析尚未被人类解读过的病理学切片)^[25]。机器学习和人类学习之间最大的区别是人类可以从少量的数据中得出复杂的关联。例如,一个蹒跚学步的孩子并不需要看到很多猫的例子,就能认出猎豹是猫科动物。一般来说,机器学习完成同样的任务需要比人类学习更多的例子。然而,另一方面,机器可以从大量的数据中进行学习。研究者完全可以使用存储在电子健康记录中的成千上万的患者病历,以及数以千亿计的数据,将机器学习模型训练到没有任何失误的程度,而人类的医生在整个职业生涯都很难看到成千上万的病例。

从更专业的角度来讲,机器学习是通过适应数据模式在算法上拟合模型的技术的总称,可以被分为监督学习、无监督学习、半监督学习。监督学习可以通过分析变量、结果及其关系来识别模型,优化准确性,比如自动拟合回归模型(包括任何形式的广义线性模型)就属于监督学习技术;相比之下,无监督学习则是利用输入数据集的固有属性来检测趋势和模式,而无需指定一系列的兴趣的结果,例如,识别观测数据中潜在协方差结构的主成分分析就是无监督的;半监督学习则是上述两种的混合体,用于预测作为目标但大多数数据缺少结果信息的情况,半监督和无监督学习通常用于数据挖掘阶段^[15]。机器学习如今已经在数据科学领域中得到了更广泛的应用,一些公共卫生研究人员也逐渐将机器学习运用到公共卫生领域。例如,无监督学习已被用于空间和时空分析,暴发确认和监测,根据临床结果识别相关患者的特征以及环境监测;利用半监督学习来构建基于社交媒体数据的不良药物反应的早期预警系统,根据智能手机的数据监测意外跌落情况,以及识别异常空气污染物以及其他应用;监督学习已被用于预测医院再入院率,结核病的传播,机动车事故中的严重伤害以及社交网站用户转向自杀意念的情况^[15]。

四、健康大数据的应用

三级预防一直都是公共卫生人员进行预防控制的经典指南,为慢性病的预防控制提供方向。而传统的三级预防方法需要耗费大量的人力物力,例如,传统的风险分级是通过非电子方式(患者问卷调查,手动图表审查和亲自评估)完成的。但是,大数据的出现极大地改变了风险评分,甚至将队列划分为更加详细的亚组开展研究。可以设想,研究者们可以收集和分析健康大数据,利用这些信息识别高风险个体,提供更有效的治疗方法,并在整个医疗保健系统中提出降低成本的行之有效的方法。中国作为目前世界上人口最多的国家,具有庞大的基本数据信息优势,大量的公共卫生大数据急需挖掘、整合和利用,海量的健康数据有待被收集整理。随着健康大数据的数据积累和方法发展,慢性病三级预防的研究思路和方法得到了拓展和创新。

1. 一级预防:慢性病领域的一级预防主要是针对慢性病危险因素的预防,防止高危人群发展为慢性病患者。结合健

康大数据,研究者可以探索慢性病相关的还未被发现的危险因素,研究危险因素行之有效的管理方法或是对其进行改进,在群体层面还可以利用人群危险因素情况对慢性病的患病风险作出预测,从多角度来实施慢性病的一级预防。公共卫生和管理人员还可以从政策层面,推进有利于开展危险因素预防的群体性防控措施,更全面、更广泛地开展一级预防。

过去由于研究样本量的限制,研究者通常只能在较小的研究对象中探索慢性病的危险因素。而随着健康大数据存量的不断积累,研究者有了更多的样本数据可以用于探索被人们所忽视的危险因素,甚至可以结合其他类型的大数据,探索更多潜藏的危险因素。例如 Xing 等^[26]利用 2016 年中国癌症登记报告中的肺癌发病数据和中国国家环境监测中心的空气检测数据,采用 Geo Detector 软件分析了中国 207 个县的男女性肺癌发病率与 PM_{2.5}、PM₁₀、SO₂、NO₂、CO 和 O₃ 的年浓度之间的关系。再比如陈可^[27]依托临床大数据中心,参考 Up To Data 临床顾问数据库所列的肺癌危险因素选取研究变量,选用 logistic 回归模型建立预测模型,结合主成分分析法筛选特征变量,对肺癌发生的危险因素进行探讨,以便更好地从危险因素介入开展肺癌的预防控制。过去的很多风险预测模型都因为缺乏数据而不尽人意,现如今,越来越多的研究者利用大量的健康大数据构建慢性病危险因素风险预测模型,为慢性病的预防提供理论依据。例如万琦等^[28]利用武汉市某社区居民的健康档案数据,结合 χ^2 检验与多因素 logistic 回归模型分析高血压的危险因素,将危险因素作为输入,血压值作为输出,建立 BP 神经网络预测模型,对血压值进行预测,为高血压预防提供依据并探索医疗健康大数据的应用模式。华中科技大学的研究团队利用结构化的医院数据包括个人属性(性别、年龄、身高体重等)、生活习惯(吸烟与否)、检查结果(血常规等)和非结构化的个人患病史及历史医嘱等文本数据,基于改进的卷积神经网络,对个体脑梗患病风险进行预测,预测准确率达到 94.8%^[29]。

就慢性病预防措施的制定而言,由于有待解决的问题和面向的对象极为复杂,很多问题远非单纯的医学技术手段就能够解决。因此,非医学领域(如社会科学、经济学、教育、伦理、司法等)中的证据反而更具借鉴价值。我们既需要来自生物医学领域的证据,也需要社会、经济、心理、政治等多学科领域的证据^[30]。因此,可以利用庞大的大数据资源为循证公共卫生决策提供真实有效的支持。其中,最广为人知的利用大数据为决策提供服务的项目应该就是全球疾病负担(GBD)研究。该项目利用来自全球的数据资源,包括 WHO、各国 CDC、民政、公安、医疗机构以及专项调查等多方数据,为政府合理分配卫生资源,做出正确的决策提供有价值的信息^[1]。但中国还没有独立的、专业的循证公共卫生决策数据库,建立一个完善的、专业的循证数据库对于我国公共卫生决策的制定十分必要和迫切。除了 GBD 这样大型、全面的项目,还有许多针对某项预防措施开展的评价研究。比如,为了预防儿童肥胖,英国设立了一个目标,到 2020 年将某些高糖产品的含糖量减少 20%,从而减少儿童糖摄入。研究者

通过模型研究发现,如果减糖计划全部实施,可以直接地降低儿童和成人的每日卡路里,4~10岁的儿童每天卡路里减少量为25(95%CI:23~26)kcal,11~18岁儿童每天卡路里减少量为25(95%CI:24~28)kcal,成年人则为19(95%CI:17~20)kcal;还可以有效地降低儿童肥胖,在接受调查的4~10岁儿童中,肥胖儿童数降低了0.6%^[31]。再比如,医院的再入院率是评价卫生系统的一项重要质量指标,对卫生保健的成本影响也很大。2011年,仅因30 d再入院就花费了美国医院超过413亿美元。德克萨斯州达拉斯利用帕克兰卫生与医院系统的电子病历数据对心力衰竭患者再次入院的风险进行预测。研究结果显示,干预后入院的患者与干预前入院的患者相比,经风险调整后的再次入院风险比相对降低了26%^[32]。这些可以为政府的卫生决策提供证据支持,也可以为政府的卫生资金分配提供依据,还可以提高卫生资源的最优化利用方案,从“关口”处实现慢性病的预防控制。

2. 二级预防:又称为“三早预防”,指的是早发现、早诊断和早治疗。大数据时代之前的二级预防主要是通过普查、筛查、定期健康检查、高危人群重点项目检查、设立专科门诊等方法开展。但耗费人力、筛查不全等问题也一直困扰着研究者。随着大数据时代的到来,研究者可以采用更便捷快速的方式发现高危人群和患者,提高诊断的精确性和快速性,更早地干预,控制疾病的进程,减少严重后遗症的出现。

利用健康大数据技术与方法可将传统的健康数据(如电子和纸质病历等)与其他来源的个人数据(如饮食、睡眠、锻炼习惯、生活方式、社交媒体和休闲、收入、教育等)联系起来进行健康监测和管理^[33~34],更全面地实现“早发现”,尽早对发现的病例进行管理。比如,可以利用可穿戴设备或是智能设备对个体的血压、血糖、心率、脉搏、热量消耗进行测量并记录在云平台上,甚至可以监测个体的生活行为规律、睡眠质量、饮食状况、体育锻炼频率、烟酒摄入量等。云平台的算法可以对这些数据进行处理,并生成个性化的健康评估报告,对于高风险的个体还可进行风险评估和预警^[35]。

不仅仅是利用移动健康技术的动态监测数据,利用已有的健康大数据也可以为“早诊断”献计献策。积水潭医院研究团队开展了前瞻性、多中心的队列研究,在研究对象知情同意的情况下,对健康体检人群进行腰椎骨密度、腹内脂肪分布和脂肪肝的QCT测量(筛查)及后续随访工作^[36]。旨在建立中国人群的腰椎骨密度、腹内脂肪和肝脏脂肪含量的正常参考值数据库,为确立中国人群骨质疏松症、肥胖和脂肪肝的QCT诊断标准提供参考依据,实现“早诊断”的突破,进而为卫生相关部门制定中国人群骨质疏松症、肥胖和脂肪肝的预防措施及政策提供基础数据。

3. 三级预防:又称为临床预防,是在疾病的临床期为了减少疾病的危害而采取的措施,主要包括对症治疗和康复治疗。在健康大数据飞速增长的背景下,症状管理、患者康复、治疗方法指南等有效方法也得到了不断的发展。

研究者可以在较大的样本量(覆盖具有广泛代表性的更大受试人群)的基础上,根据患者的实际病情和意愿非随机

地选择治疗措施,开展长期评价,关注有研究意义的结局变量,以进一步评价干预措施的外部有效性和安全性,为随机对照试验范围之外的研究提供论证依据^[37]。例如以电子健康档案(electronic health records, EHRs)和电子病历(electronic medical records, EMRs)为主要研究数据,辅以其他相关数据,在较大的人群范围内,通过对数据的研究,探索科学问题、评估患者健康状况及诊疗过程、评估防治结局、评估患者预后与预测、支持医疗决策的制定^[25]。如Pobiruchin等^[38]采用临床癌症登记处的乳腺癌患者的记录,构建真实世界的乳腺癌参考模型。该模型更详细地反映出癌症转移的概率,特别是无病生存期和复发期,可以更好地指导医生进行疾病管理^[39]。还有研究者结合个人基因谱和完整病史数据,将健康危险因素进行关联比对分析,跟踪病程进展、判断短期风险和长期预后,获得比临时就诊更准确的信息,从而进行更有效、更个性化的临床干预和健康指导^[8],在疾病的早期就开展“早治疗”,避免产生更大的疾病伤害。又如宜昌市健康管理大数据中心以网格化人口数据库为基础,采用大数据、云计算、物联网等新技术,打通公安、人社、卫生、药监、环保、安监等多部门,实现数据采集共享、互联互通^[40]。凭借这个大数据平台,截止2018年底,宜昌市城区已发现并全过程管理97 687例高血压患者、24 352例糖尿病患者、14 350例卒中患者和24 530例各类肿瘤患者等。为患者的定期科学健康管理提供了技术支持和理论指导,能够更有效地控制病情,减少并发症。

五、面临的挑战

尽管大数据如今的应用研发蓬勃发展,但仍然面临着诸如关键技术突破、数据管理和共享、数据质量和伦理学等方面挑战,制约了数据价值的有效发挥^[41]。

1. 数据的技术问题:现如今,大数据来源繁多,但当前的电子健康档案和繁多的疾病资料并没有按照结构化的形式存储,大多数的健康数据几乎都以片段式的结构存储,包括半结构数据和非结构化数据。这为数据挖掘、数据分析和数据管理带来诸多不便^[42]。而且,信息记录残缺,时效性差,描述不规范,来自网络的数据更是杂乱无章,信息噪音较大^[43]。为了方便大数据的应用和分析,需要解决数据的结构化问题、质量问题,并且开发可以将各个来源的大数据进行有效整合的技术。

2. 数据管理和共享:除上述技术挑战外,数据信息“孤岛”问题也普遍存在^[8]。由于数据结构不同以及对于数据安全的担忧等多个问题,各个平台之间难以架起互通互利的数据共享桥梁。2015年国务院通过《关于促进大数据发展的行动纲要》(《纲要》),在这个《纲要》中指出,政府应当推动公共数据互联共享、消除信息孤岛,避免重复建设和数据“打架”,整合各级平台。我国数据共享尚处于起步阶段,国家和各省市卫生健康委员会网站的“政务公开”栏目,公开了医疗机构数、诊疗费用、诊疗人次、出院人数等统计数据以及统计年报、月报、公报等数据资源;中国疾病预防控制中心、国家统计局等也公布了部分健康医疗数据^[44]。但完善数据元、核心

元数据、数据集、数据敏感程度评估、数据开放风险评估、数据采集接口等标准体系的确定,以及安全和隐私问题是我国健康医疗大数据共享开放的主要障碍之一^[45]。因此需要从技术和管理相结合的视角完善法律体系,加强数据安全与隐私管理和技术标准规范约束,提升安全与隐私泄露风险评估和国家商用密码的应用,提高社会数据安全素养,从各方面协同对应数据安全与隐私保护问题,从数据安全治理的角度解决此问题^[45]。

3. 数据的质量问题:由于各机构信息系统和标准的不一致,不同来源的数据资源格式不同,存储标准不一,同一来源渠道的资源在不同时期也会由于人员、软硬件环境等不同而存在差异^[2]。因此,在健康大数据实际使用中存在着数据可用性低、数据质量差等问题。而这可能是目前限制利用这些数据的重要障碍之一。而且健康大数据中还有许多“脏数据”,即虚假数据,对这些数据的分析将会得出不正确的结论,导致错误的预测结果,对健康管理、循证结果产生严重的负面影响^[46]。因此,甄别健康大数据的“误差”尤为重要,应当认真检查数据的真实性,医疗数据使用时要与临床经验和实际情况相结合。

4. 伦理与隐私:随着科技技术的发展,数据量的激增和数据的可用性已经引起了公众的关注,而相关的担忧也必然会随之增长^[47]。有研究者据此提出3个关于大数据隐私的关键问题^[48]:第一,无意披露个人身份信息的风险,例如,使用网络在线工具而泄露隐私;第二,数据维度的增加,使得更加难以确定数据集是否被取消识别,也更难防止个人识别信息的披露;第三,面对可能改变传统规范的新技术,很多原本可以得到很好保护的隐私面临威胁,例如,GPS、无人机、社交媒体等。因此需要制定相关政策法规,分辨出哪些数据属于隐私数据,哪些数据可以共享和利用,明确隐私保护对象。同时应加大力度完善技术规范,如利用数据脱敏、去标签化等手段对数据进行处理,尽可能保护数据隐私^[2]。

健康大数据具有极大的潜能和研究价值,但同时也面临着巨大的来自技术方面和设计方面的挑战。从繁多复杂的大数据中快速地分析获得有价值的信息,为慢性病的预防控制提供良方良策,这是一项长久的工作,它不仅需要政府的鼎力支持,精心设计,也需要技术人员在数据收集、存储、挖掘等方面取得可靠的进展,还需要大批的科研人员的钻研创新。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

- [1] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017 [J]. Lancet, 2018, 392 (10159) : 1789–1858. DOI: 10.1016/S0140-6736(18)32279-7.
- [2] 张翼鹏,黄竹青,陈敏.公共卫生大数据应用模式探讨[J].中国数字医学,2019,14(1):33-35. DOI: 10.3969/j.issn.1673-7571.2019.01.010.
- Zhang YP, Huang ZQ, Chen M. Big data application model in public Health[J]. China Digital Med, 2019, 14(1):33-35. DOI: 10.3969/j.issn.1673-7571.2019.01.010.
- [3] 孟群.促进健康医疗大数据应用保障“健康中国2030”建设[J].中国卫生信息管理杂志,2016,13(6):539.
- Meng Q. Promote health care big data application guarantee the construction of “Healthy China 2030”[J]. Chin J Health Inform Manag, 2016, 13(6):539.
- [4] Biro S, Williamson T, Leggett JA, et al. Utility of linking primary care electronic medical records with Canadian census data to study the determinants of chronic disease: an example based on socioeconomic status and obesity [J]. BMC Med Inform Decis Mak, 2016, 16:32. DOI: 10.1186/s12911-016-0272-9.
- [5] Hemingway H, Feder GS, Fitzpatrick NK, et al. Using nationwide ‘big data’ from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the Clinical disease research using Linked Bespoke studies and Electronic health Records (CALIBER) programme [J]. Programme Grants Appl Res, 2017, 5 (4) : 1-330. DOI: 10.3310/pgfar05040.
- [6] Li HL, Lin HB, Zhao HY, et al. Statins use and risk of new-onset diabetes in hypertensive patients: a population-based retrospective cohort study in Yinzhou district, Ningbo city, People’s Republic of China [J]. Ther Clin Risk Manag, 2018, 14: 823-832. DOI: 10.2147/TCRM.S158850.
- [7] Zhong R, Wu YS, Cai YP, et al. Forecasting hand, foot, and mouth disease in Shenzhen based on daily level clinical data and multiple environmental factors [J]. Biosci Trends, 2018, 12 (5) : 450-455. DOI: 10.5582/bst.2018.01126.
- [8] 孟润堂,罗艺,宇传华,等.健康大数据在公共卫生领域中的应用与挑战[J].中国全科医学,2015,18(35):4388-4392. DOI: 10.3969/j.issn.1007-9572.2015.35.029.
- Meng RT, Luo Y, Yu CH, et al. Application and challenges of healthy big data in the field of public health [J]. Chin General Pract, 2015, 18(35) : 4388-4392. DOI: 10.3969/j.issn.1007-9572.2015.35.029.
- [9] Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care [J]. Chest, 2018, 154 (5) : 1239-1248. DOI: 10.1016/j.chest.2018.04.037.
- [10] 赵自雄,史倩楠,马家奇.公共卫生大数据应用实例与发展建议[J].中国卫生信息管理杂志,2017,14(5):655-659. DOI: 10.3969/j.issn.1672-5166.2017.05.04.
- Zhao ZX, Shi QN, Ma JQ. Application examples and development recommendations of big data in public health [J]. Chin J Health Inform Manag, 2017, 14 (5) : 655-659. DOI: 10.3969/j.issn.1672-5166.2017.05.04.
- [11] Zhang XZ, Pérez-Stable EJ, Bourne PE, et al. Big data science: opportunities and challenges to address minority health and health disparities in the 21st century [J]. Ethn Dis, 2017, 27 (2) : 95-106. DOI: 10.18865/ed.27.2.95.
- [12] 竺忠林,王雅洁,陈娟,等.健康大数据在医疗卫生领域中的应用及挑战[J].海南医学,2017,28(2):173-176. DOI: 10.3969/j.issn.1003-6350.2017.02.001.
- Mou ZL, Wang YJ, Chen J, et al. Application and challenges of medical health big data in the field of health care [J]. Hainan Med J, 2017, 28 (2) : 173-176. DOI: 10.3969/j.issn.1003-6350.2017.02.001.
- [13] 王黎洲.健康大数据在公共卫生领域中的应用研究[J].中国卫生标准管理,2016,7(9):1-2. DOI: 10.3969/j.issn.1674-9316.2016.09.001.
- Wang LZ. Research on the application of health big data in the field of public health [J]. China Health Stand Manag, 2016, 7 (9):1-2. DOI: 10.3969/j.issn.1674-9316.2016.09.001.
- [14] 段金宁.“互联网+”医疗环境下的健康医疗大数据应用[J].中华医学图书情报杂志,2018,27 (6) : 49-53. DOI: 10.3969/j.issn.1671-3982.2018.06.008.
- Duan JN. Application of healthcare big data in “Internet +” healthcare [J]. Chin J Med Lib Inf Sci, 2018, 27 (6) : 49-53. DOI: 10.3969/j.issn.1671-3982.2018.06.008.
- [15] Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy [J]. Ann Rev Public Health, 2018, 39:95-112. DOI: 10.1146/annurev-publichealth-040617-014208.
- [16] Rajkomar A, Dean J, Kohane I. Machine learning in medicine [J]. N Engl J Med, 2019, 380 (14) : 1347-1358. DOI: 10.1056/NEJMra1814259.
- [17] Khan N, Yaqoob I, Hashem IAT, et al. Big data: survey, technologies, opportunities, and challenges [J]. Scientific World J, 2014, 2014:712826. DOI: 10.1155/2014/712826.
- [18] Lavrač N, Bohanec M, Pur A, et al. Data mining and visualization for decision support and modeling of public

- health-care resources [J]. *J Biomed Inform*, 2007, 40 (4) : 438–447. DOI:10.1016/j.jbi.2006.10.003.
- [19] 史倩楠,马家奇.公共卫生大数据分析方法与应用方向[J].中国数字医学,2016,11(2):10–12. DOI:10.3969/j.issn.1673–7571.2016.02.003.
Shi QN, Ma JQ. Big data analytics and application in public health [J]. *China Digital Med*, 2016, 11 (2) : 10–12. DOI: 10.3969/j.issn.1673–7571.2016.02.003.
- [20] 李照东,吴建林.基于关联规则的电子病历数据挖掘应用研究[J].江苏科技信息,2018(8):56–59.
Li ZD, Wu JL. Application of electronic medical record data mining based on association rules [J]. *Jiangsu Sci Technol Inf*, 2018(8):56–59.
- [21] 章涛,官海滨,李傅冬,等.应用Elman神经网络建立流感样病例预测模型[J].预防医学,2019,31(2):113–118. DOI:10.19485/j.cnki.issn2096–5087.2019.02.002.
Zhang T, Guan HB, Li FD, et al. Modeling of influenza-like illness prediction based on Elman neural network [J]. *J Prev Med*, 2019, 31(2):113–118. DOI:10.19485/j.cnki.issn2096–5087.2019.02.002.
- [22] 王董磊,张雪峰,王笑辰,等.基于动力学-决策树模型的狂犬病免疫策略成本效用分析[J].中华预防医学杂志,2019,53(8):804–810. DOI:10.3760/cma.j.issn.0253–9624.2019.08.002.
Wang DL, Zhang XF, Wang XC, et al. Cost-effectiveness analysis of rabies immunization strategy based on dynamic-decision tree model [J]. *Chin J Prev Med*, 2019, 53(8):804–810. DOI:10.3760/cma.j.issn.0253–9624.2019.08.002.
- [23] 杨广锋,胡晓娟,张湘星.医疗大数据应用和技术若干问题探讨[J].现代信息科技,2019,3(9):106–108. DOI:10.3969/j.issn.2096–4706.2019.09.041.
Yang CF, Hu XJ, Zhang XX. Discussion on application and technology of medical big data [J]. *Mod Inf Technol*, 2019, 3(9): 106–108. DOI:10.3969/j.issn.2096–4706.2019.09.041.
- [24] 杨磊.大数据的发展及数据仓库的融合应用[J].数字技术与应用,2019,37(6):62, 64. DOI:10.19695/j.cnki.cn12–1369.2019.06.34
Yang L. The development of big data and the fusion application of data warehouse [J]. *Digital Technol Appl*, 2019, 37 (6) : 62, 64. DOI:10.19695/j.cnki.cn12–1369.2019.06.34
- [25] 孙鑫,谭婧,唐立,等.重新认识真实世界研究[J].中国循证医学杂志,2017,17(2):126–130. DOI: 10.7507/1672–2531.201701088.
Sun X, Tan J, Tang L, et al. Revisiting real-world study [J]. *Chin J Evid-Based Med*, 2017, 17 (2) : 126–130. DOI: 10.7507/1672–2531.201701088.
- [26] Xing DF, Xu CD, Liao XY, et al. Spatial association between outdoor air pollution and lung cancer incidence in China [J]. *BMC Public Health*, 2019, 19 (1) : 1377. DOI: 10.1186/s12889–019–7740–y.
- [27] 陈可.基于电子病历大数据分析的疾病预测建模[J].中国数字医学,2018,13(3):16–18. DOI:10.3969/j.issn.1673–7571.2018.03.005.
Chen K. Construction of disease prediction model based on EMR big data analytics [J]. *China Digital Med*, 2018, 13 (3) : 16–18. DOI:10.3969/j.issn.1673–7571.2018.03.005.
- [28] 万琦,王威,黄薇,等.基于武汉市某社区居民电子健康档案的高血压预测模型[J].现代预防医学,2018,45(6):1030–1033.
Wan Q, Wang W, Huang W, et al. Prediction model of hypertension based on electronic health records of a community resident in Wuhan [J]. *Mod Prev Med*, 2018, 45(6):1030–1033.
- [29] Lin YK, Chen H, Brown RA, et al. Healthcare predictive analytics for risk profiling in chronic care: a Bayesian multitask learning approach [J]. *MIS Quart*, 2017, 41(2):473–495. DOI: 10.25300/MISQ/2017/41.2.07.
- [30] 李立明,吕筠.关注循证公共卫生决策[J].中华流行病学杂志,2006,27(1):1–4. DOI:10.3760/j.issn:0254–6450.2006.01.001.
Li LM, Lv J. Pay close attention to the evidence-based public health policy [J]. *Chin J Epidemiol*, 2006, 27 (1) : 1–4. DOI: 10.3760/j.issn:0254–6450.2006.01.001.
- [31] Amies-Cull B, Briggs ADM, Scarborough P. Estimating the potential impact of the UK government's sugar reduction programme on child and adult health: modelling study [J]. *BMJ*, 2019, 365:i11417. DOI:10.1136/bmj.i11417.
- [32] Bhardwaj N, Wodajo B, Spano A, et al. The impact of big data on chronic disease management [J]. *Health Care Manager*, 2018, 37 (1):90–98. DOI:10.1097/HCM.0000000000000194.
- [33] Kuziemsky CE, Monkmann H, Petersen C, et al. Big Data in healthcare—defining the digital persona through user contexts from the micro to the macro: contribution of the IMIA organizational and social issues WG [J]. *Yearb Med Inform*, 2014, 23(1):82–89. DOI:10.15265/IY-2014-0014.
- [34] Bellazzi R. Big data and biomedical informatics: a challenging opportunity [J]. *Yearb Med Inform*, 2014, 9 (1) : 8–13. DOI: 10.15265/IY-2014-0024.
- [35] 徐明珍,汪栋材,吴海滨,等.大数据背景下慢性病健康管理系统的构建研究[J].工程技术研究,2018(2):250–251. DOI:10.3969/j.issn.1671–3818.2018.02.119.
Xu MZ, Wang DC, Wu HB, et al. Research on the construction of chronic disease health management system under the background of big data [J]. *Metall Collect*, 2018 (2) : 250–251. DOI:10.3969/j.issn.1671–3818.2018.02.119.
- [36] 过哲,付晓霞,唐雪,等.中国健康定量CT大数据项目研究方案[J].中华健康管理学杂志,2018,12(6):510–513. DOI: 10.3760/cma.j.issn.1674–0815.2018.06.003.
Guo Z, Fu XX, Tang X, et al. Health big data protocol for quantitative computed tomography in China (China Biobank) [J]. *Chin J Health Manag*, 2018, 12(6):510–513. DOI:10.3760/cma.j.issn.1674–0815.2018.06.003.
- [37] 梁远波,吴越,郑景伟.“真实世界”研究的产生背景、概念、方法及其在眼科的应用[J].中华眼视光学与视觉科学杂志,2013,15(12):756–759. DOI: 10.3760/cma.j.issn.1674–845X.2013.12.014.
Liang YB, Wu Y, Zheng JW. Background, conceptions, methodology of the Real-World Study and its application in ophthalmology [J]. *Chin J Opton Ophthalmol Vis Sci*, 2013, 15 (12) : 756–759. DOI:10.3760/cma.j.issn.1674–845X.2013.12.014.
- [38] Pobiruchin M, Bochum S, Martens UM, et al. A method for using real world data in breast cancer modeling [J]. *J Biomed Inform*, 2016, 60:385–394. DOI:10.1016/j.jbi.2016.01.017.
- [39] Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare [J]. *IEEE J Biomed Health Inform*, 2015, 19(4):1209–1215. DOI:10.1109/JBHI.2015.2406883.
- [40] 刘建华,张培,徐承中,等.健康管理大数据构建与实践[J].中华流行病学杂志,2019,40(2):227–230. DOI:10.3760/cma.j.issn.0254–6450.2019.02.019.
Liu JH, Zhang P, Xu CZ, et al. Construction and use of big data for health management [J]. *Chin J Epidemiol*, 2019, 40 (2) : 227–230. DOI:10.3760/cma.j.issn.0254–6450.2019.02.019.
- [41] 王超,郭振,蔡云飞.大数据时代下企业大数据应用管理体系的探讨[J].电脑知识与技术,2017,13(27):246–248. DOI: 10.14004/j.cnki.ckt.2017.3160.
Wang C, Guo Z, Cai YF. Discussion on application management system of enterprise big data in the era of big data [J]. *Computer Know Technol*, 2017, 13 (27) : 246–248. DOI: 10.14004/j.cnki.ckt.2017.3160.
- [42] 范卢明,梁桂仙.大数据在健康管理中的应用研究进展[J].中国全科医学,2016,19(31):3786–3789. DOI:10.3969/j.issn.1007–9572.2016.31.005.
Fan LM, Liang GX. Research progress of application of big data in health management [J]. *Chin General Pract*, 2016, 19 (31) : 3786–3789. DOI:10.3969/j.issn.1007–9572.2016.31.005.
- [43] 朱蕊,彭龔.医疗大数据的应用[J].中国西部科技,2015,14(5):95–97. DOI:10.3969/j.issn.1671–6396.2015.05.040.
Zhu R, Peng Y. The application of medical big data [J]. *Sci Technol West China*,2015,14(5):95–97. DOI:10.3969/j.issn.1671–6396.2015.05.040.
- [44] 卞海燕,陈敏.健康医疗大数据开放管理探讨[J].中华医院管理杂志,2019,35(8):660–663. DOI:10.3760/cma.j.issn.1000–6672.2019.08.011.
Mou HY, Chen M. Discussions on the management of healthcare big data open access [J]. *China Hosp Admin*, 2019, 35 (8) : 660–663. DOI:10.3760/cma.j.issn.1000–6672.2019.08.011.
- [45] 丁红发,孟秋晴,王祥,等.面向数据生命周期的政府数据开放的数据安全与隐私保护对策分析[J].情报杂志,2019,38(7):151–159. DOI:10.3969/j.issn.1002–1965.2019.07.023.
Ding HF, Meng QQ, Wang X, et al. Countermeasure analysis for data security and privacy issues of government data opening in the view of data life-cycle [J]. *J Intellig*, 2019, 38 (7) : 151–159. DOI:10.3969/j.issn.1002–1965.2019.07.023.
- [46] Webster PC. Big data's dirty secret [J]. *CMAJ*, 2013, 185 (11) : e509–510. DOI:10.1503/cmaj.109–4516.
- [47] Mai JE. Three models of privacy. New perspectives on informational privacy [J]. *Nord Inform*, 2016, 37 Suppl 1: 171–175. DOI:10.1515/nor–2016–0031.
- [48] Bader MDM, Mooney SJ, Rundle AG. Protecting personally identifiable information when using online geographic tools for public health research [J]. *Am J Public Health*, 2016, 106 (2) : 206–208. DOI:10.2105/AJPH.2015.302951.