DOI:10.3969/j.issn.1672-5565.201705005

公共卫生大数据研究进展——生物信息的新领域

马天有[1],胡 曦[2], 王丽娜[2],杜建强[2],吴晓明[2*]
(1. 环境与疾病相关基因教育部重点实验室(西安交通大学),西安 710061;
2. 生物医学信息工程教育部重点实验室(西安交通大学),西安 710049)

摘 要:大数据时代的公共卫生面临新的机遇和挑战。为了推动公共卫生大数据的应用,准确把握其内涵,开发针对性的解决方案,达到改善人们健康状况的目的,基于此对公共卫生大数据的现状进行了分析和论证。研究表明:通过对多个不同来源公共卫生数据进行收集和整理,能够形成公共卫生大数据,通过深入挖掘和分析,能够获取重大疾病影响因素、流行病的传播规律等信息,帮助医疗卫生人员和相关机构进行预测和评估,以便采取有效的管理手段和措施,保护人民健康,减少医疗花费。发现通过同生物信息技术相结合,公共卫生数据的获取、管理、分析、安全和应用方面都会有很大的发展空间。认为计算机技术的进一步应用,针对性的大数据挖掘方法开发,以及新型公共卫生人才培养,是发展这一领域的关键因素。

关键词:公共卫生;大数据;生物信息;数据挖掘;互联网
中图分类号:R181 **文献标志码**:A **文章编号**:1672-5565(2017)04-255-08

New field in bioinformatics——progress of big data research in public health

MA Tianyou[1], HU Xi[2], WANG Lina[2], DU Jianqing[2], WU Xiaoming[2*]
(1. *Key Laboratory of Environment and Diseases Related Genes of Ministry of Education* (*Xi'an Jiaotong University*), *Xi'an* 710061, *China*;
2. *Key Laboratory of Biomedical Information Engineering of Ministry of Education* (*Xi'an Jiaotong University*), *Xi'an* 710049, *China*)

Abstract: The public health research faces new challenges and opportunities in the big data era. To promote the application of big data in public health, to understand its basics connotation precisely and to achieve the goal of improving people's health problem by developing solutions to mine information from the big data, this paper presents the current situation of big data in public health field. It shows that big data will form by collecting and tidying public health related data from different sources. Through in-depth mining and analysis, information relating to disease spread and health threaten would be uncovered and evaluated. Accordingly, measures and suggestions could be made to prevent threatens, to protect people's health, and to reduce total medical costs. It is also found that by combining with bioinformatics technology, public health research techniques such as data acquisition, system management, information security, and application would have a large room to grow. In conclusion, the application of computer technology, the development of big data mining method, and the cultivation of public health related personnel, are effective factors in the development of this field.

Keywords: Public Health; Big data; Bioinformatics; Data mining; Internet

目前健康相关的检测和测试手段,产生了大量数据,这些数据包括来自医院的门诊和临床数据、家庭小型便携式设备的检测监护数据、医疗保险机构的就医数据,可穿戴设备产生的个人健康数据、以及来自公共平台的人口、微生物分布、食品保健、产品销售等信息。这些信息能从不同角度对公共卫生相关情况进行呈现。通过利用并开发合适的数据处理和挖掘方法,能发现公共卫生数据中隐含的信息,并形成指导改善公共健康的方案和措施。但是由于信息的多元化和不确定性,此类数据如何进行有效利

收稿日期:2017-05-19;**修回日期**:2017-07-08.
基金项目:陕西省科技计划项目(2011K12-02-04, 2011K12-04-06).
作者简介:马天有,男,博士,研究方向:公共卫生; E-mail: maty@mail.xjtu.edu.cn.
* **通信作者**:吴晓明,男,副教授,研究方向:生物信息技术;E-mail:wxm@mail.xjtu.edu.cn.

用,需要政策、技术、资金、算法、数据管理等多方面的支撑。

通讯网络不断深入到日常生活,将成为获取和传播公共卫生数据的重要手段。对其进行合理利用,可将信息采集功能拓展到更广泛的领域,也有助于提高传染病、突发事件监测的准确性,以便科学合理地实现快速响应,降低疾病和公共卫生事件的危害。基因数据和健康数据涉及生命活动机理,对其进行挖掘和分析,也可以提供更为准确的风险评估及个体化干预措施,人们可就此改变不良生活习惯,减少危险因素[1]。进行公共卫生数据分析将在健康领域发挥重要作用。本文对公共卫生大数据研究的方法、技术和前景进行探讨,认为需要从政策、人才、硬件等方面形成支持,同时对数据进行收集、管理、分析和挖掘,最终形成个人和社会的受益。

1 公共卫生数据来源广泛

用于公共卫生研究的数据有非常广泛的来源,它们互相补充,相互支撑,能体现群体性健康问题的各种特征,数据的种类可以包括以下几方面:

1)医疗数据。医疗机构拥有患者个人的多种信息,其中的临床数据是和个人健康密切相关的信息。当作为整体进行考察时,能够体现同大规模公共卫生事件相关的信息。随着中国推进分级诊疗和家庭医生签约服务,家庭医生更能够对患者进行健康监测,形成健康数据,会成为公共卫生数据的重要源头。

2)家庭护理和便携式设备检测数据。家用智能健康测量装置均可产生和记录健康相关数据。一些产品已经面市,包括智能体重秤、蓝牙血糖仪、电子血压计等。智能手环、计步器、专门测量呼吸的运动背心等,也可产生大量健康数据。如 WellDoc 公司研发的基于手机 App 和云端大数据的糖尿病管理平台,是获得美国食品和药品管理局批准的手机应用,用户可以通过手机实时记录、存储和利用糖尿病数据。通过进行实时挖掘分析,可为患者提供个性化反馈,指导患者进行改变生活方式,并为医生的诊疗提供有效依据[1-2]。这些数据的特点是数据量大、种类繁多,但准确性较差,需进行有效的校准和过滤方可使用。

3)地理信息数据。由于公共卫生数据大都具有空间属性,进行大数据分析时也常结合地理信息系统(GIS)来分析研究其空间特征和规律[3]。通过结合地理位置、行政区域、气象条件等,数据的空间特点可以进一步体现。

4)生物医学数据库和政府基础平台。互联网的各类公共生物数据库提供了有关生物分子、微生物分类等的详细信息[4]。中国最新建设的“国家人口与健康科学数据共享平台”(http://www.ncmi.cn/1),也已经包含 237 个数据集,数据量达到 49.1 TB,覆盖包括生物医学、基础医学、临床、公共卫生、中医药学、药学、人口与生殖健康七大类,将带动生物医学数据资源整合与共享,为实现健康中国 2030 年的战略目标发挥作用。

5)其他数据。气象、舆情、疫情、农作物和食品安全等数据,均可用于公共卫生研究。未来,数据的种类和数量将会继续增加。事实上,所有可用于进行公共卫生状况分析的数据,都应该被考虑,并被广泛收集,从而形成全面的数据支撑。但是,这些数据存在非常高的异质性,数据中有价值信息少,含金量不高,需要采用合适的数据管理和分析方法才能够达到对数据的有效利用。

2 数据的复杂性和有效管理面临的挑战

当数据不断的被收集整理之后,随之而来产生了对大数据管理的软硬件系统和管理模式的需求,而数据的复杂性为有效解决这个问题提出新的挑战。图 1 是进行公共卫生大数据分析研究的典型框架。公共卫生大数据来源广泛,其种类和格式也在随技术进步不断变化,数据规模均不相同,需要开发相应的存取技术和数据管理方式。传统数据库 MySQL,新型非关系化数据库 Mongodb,内存数据库 TimesTen 等,能够在一定程度上对数据进行管理,但当数据规模和种类更大时,需要分布式数据库来实现。目前的技术能够实现数据的管理,例如支持淘宝的 Oceanbase,能够管理数百 T 的数据,但需要多台服务器,成本高昂。对数据管理的瓶颈是数据的异质性,不同类型的数据需要有针对性的过滤、导入、检索模块,通过合适的接口,把数据转换成为标准的形式,对软件开发提出了很高的要求。

3 生物信息数据库提供广泛支持

生物领域有多种实验数据,已经有完备的数据库系统进行数据管理,也提供给公众进行免费访问和数据检索,它们能对公共卫生数据的管理和应用提供借鉴。例如,许多生物数据库提供数据分析功能。利用 NCBI 的 blast(ncbi.nlm.nih.gov/blast/)能够进行序列比对和检索,Lynx 等数据库提供富集分析等功能,Reactome 提供网络可视化功能[5],UCSC

Xena(http://xena.ucsc.edu/)也提供针对多种临床数据结合基因数据的分析方法。广泛的生物数据库形成了解读生命活动规律的知识库,对于公共卫生数据分析提供的重要支撑。同时,生物数据库也正在走向广泛集成和存贮、分析并重的方向,其技术手段和分析流程也为公共卫生数据分析提供借鉴。

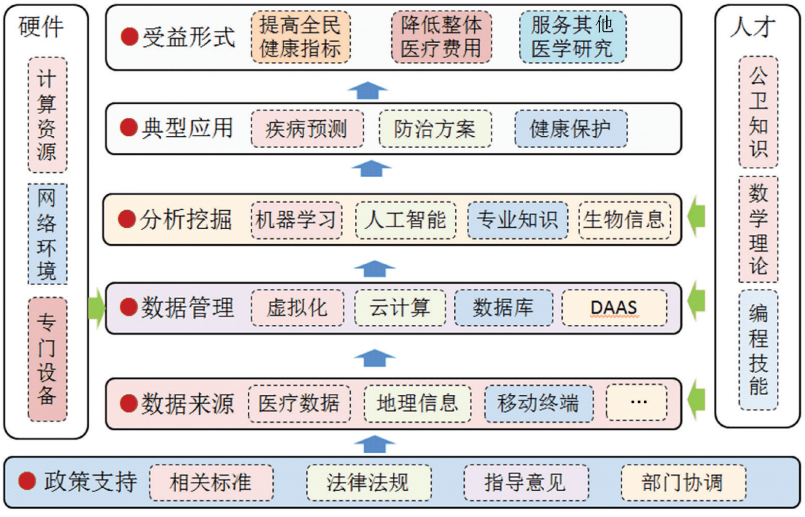

图 1　公共卫生大数据研究框架及其应用

Fig.1　Framework of public health big data research and its application

4　数据挖掘方法不断创新

通过数据挖掘能获取数据中和公共卫生相关的信息,而分析方法的选择对于获取有效结果非常关键。传统的统计学手段将继续发挥重要作用,而基于机器学习和人工智能的方法,能够包容多种不同的数据形式,并形成对数据的深度分析。基于神经网络、HMM 模型、动态规划、贝叶斯推断、随机森林的分析方法,也普遍应用于医疗、卫生数据的分析[6-7]。这也对软件开发、计算资源的使用提出了更高的要求。

网络是对复杂系统建模的基本工具[8]。公共卫生中的数据可以通过网络进行表示,利用网络模块识别技术,可找出模块之间的关联,并发现普遍存在于复杂系统中的高阶信息组织和协调方式,非常适合对流行病传播等公共卫生问题进行描述。

公共卫生数据往往涉及不同来源、不同类型的数据,而异构图(heterogeneous graph)、贝叶斯网络(Bayesian network)等可以表示不同信息之间的联系。通过图的挖掘、聚类、排序、分割、可视化,可以对不同类型的公共卫生数据进行融合分析,获取传统方法难以得到结果[9]。

研究表明,通过大数据分析,发现传统体检数据包含同心血管疾病,死亡率相关信息,而智能工具可作为评估总体健康状况的手段[10];利用谷歌趋势搜索(trends.google.com),根据各地区感染病例情况建立动态预测模型,可以对 zika 病毒的传播进行预测和防范[11]。

在 2014 年的 Ebola 疫情控制中,专家利用流行病学数据建立了相关模型,预测了 Ebola 疫情的严重后果[12-13]。同时,人工智能、机器学习被证明非常具有潜力。谷歌的深度网平台 TensorFlow 已在医学影像识别和疾病判断方面取得很好的成果,甚至能够辅助临床诊断[14],在多个研究中发挥作用。通过设定场景模式,新的公共卫生大数据分析方法将借助人工智能平台而出现。

5　计算平台提供支持

公共卫生大数据分析也需要大量的计算资源,可从 3 个层次进行配置。

1)传统的以云计算、分布存储和高性能计算为主体的计算平台。这种方式通过增加硬件,以及软件虚拟化的技术,管理大规模计算资源,提供分析和计算服务。目前的大数据处理平台和工具中,MapReduce 提供计算的分解和整合,Hadoop 提供可扩展的平台支持,HDFS 技术提供分布式的大数据存储,Hive 提供数据库的分布式管理和检索。此种平台的优点是适用性好、技术成熟、软件配置灵活。缺点是成本高、能耗高。

2)专门硬件的使用。基于 GPU、FPGA 的专门

硬件,在一个芯片上可以部署上千计算单元或逻辑电路,能够大大加速计算过程,对于需要进行反复迭代、包含大量简单操作的计算而言,是最佳的选择,其优点是性能高、成本低、能耗低,缺点是开发难度大、适用范围窄,适合对特定问题的解决。目前已经有专门芯片,进行癫痫的及时预测[15]。在生物信息领域,基于 FPGA 的技术在序列比对方面,也显示出功耗低、速度快的特点[16]。更多的专门芯片也将会有越来越多的应用于医疗和公共卫生方面。

3)超级计算及下一代计算技术。大规模的并行计算能够成倍的提高计算速度,实现海量数据存贮,使大规模的数据处理成为可能。中国开发的天河二号由 16 000 个节点组成,每个节点有 2 颗 Xeon 处理器和 3 个 Xeon Phi 处理器。持续计算速度每秒 3.39 亿亿次双精度浮点运算。2016 年 6 月,使用中国自主芯片"SW26010"制造的"神威太湖之光",包含 40 960 个处理器,浮点运算速度为每秒 9.3 亿亿次,取代天河二号登上超算榜首。这些计算能力足以同时处理大量数据。在超算平台,许多难以求解的问题都可以得到快速处理,通过并行的方式,实现高复杂度问题的求解。

6 网络技术推动应用的普及

移动互联网目前已经有很大的覆盖面,骨干互联网也已经实现高速的互联互通,为多种公共卫生大数据的收集提供技术支持。借助物联网技术,各种便携式终端、嵌入式设备借助低功耗通讯技术,可以实现地理区域大跨度、长时间的数据采集和获取。

大数据分析可从两个维度实现。一是计算机的角度,利用计算能力和人工智能,进行数据分析处理。另一维度以人作为分析主体,进行人机交互,将人所具备的认知能力融入分析过程中[17]。此时,数据的交互可视化尤为重要。基于网络,可通过浏览器收集用户分析需求,利用后台服务器实现分析结果,然后通过可视化界面显示给用户,实现交互处理和分析,大大提高获取分析结果的效率。HTML5 包含有丰富的网页对象表示方式,获得了广泛的支持,为网络应用的开发提供了很好的支持。PHP(hypertext preprocessor)超文本处理程序,能实现数据库处理,响应用户请求。AJAX(asynchronous JavaScript and XML)能实现网页和服务器之间的交互操作,并能达到实时响应的效果。借助 D3.js,vis.js,CartoDB 等工具能够增强数据的可视化效果,可以形成的显示方式包括层次数据、空间映射、时变数据、地理信息、空间标量等,凸显分析结果。R、Python 能够实现多种数据的统计分析,是服务器端分析程序的最佳选择,这些技术的综合运行,能形成基于网络的可视化。

生物信息领域的可借鉴平台是 Galaxy,它实现基于网络的数据分析过程人机交互,大大方便了数据分析流程[18]。Biomart 数据库平台本身提供数据分析服务,同时能够连接多个后台数据库,提供隔离的访问;Ensemble 包含基因组信息也提供了互联网服务器,利用标准 SQL 语句实现数据访问。对于公共卫生数据,此类数据管理方法仍旧可行。通过开发公共卫生数据处理模块和网络分析接口,用户可以自行选择分析模块,组建分析流程,实现交互式的数据分析,将会极大的推进公共卫生大数据的分析和应用。

7 有效的专业知识发挥重要作用

数据之间的联系,有些通过专业知识可被推理和演绎,揭示隐含信息,因此,把已知知识融合到数据分析中非常有效,但也具有相当大的挑战。针对公共卫生领域,宏观的疾病流行程度、群体健康状况,针对个体的体检指标、精神、心理、慢性病数据等,都需要用专业的术语和特定的统计方式进行表示。在生物信息领域,许多知识已经整理和校对,形成基础知识库,并利用生物信息的方法进行表示和处理。例如 KEGG 和 Reactome 都包含代谢网络等分子互作信息,利用网络的形式表示分子之间联系的生物学知识;Uniprot 包含有已知蛋白质的修饰、结构域知识,这些基础为生命科学研究提供了重要支撑。作为类比,公共卫生领域有药品分类信息等,ICD10 分类系统(http://www.icd10data.com/),MeSH 医学主题词系统(https://www.nlm.nih.gov/mesh/),但此类知识库还非常少。当分析结果能够同专业知识库结合时,才能达到对公共卫生信息的最佳应用,因此构建公共卫生知识库将是重要的发展领域。

8 网络安全和防护不可或缺

2017 年 5 月,勒索病毒 WannaCry 利用微软 SMB 服务漏洞(MS17-010)开始在全球大范围传播,充分说明网络安全的重要性。云平台相对于个人计算机,安全性有非常大的提高,但由于操作系统的漏洞不能被全部检测出来,因此未知漏洞的防范,已知漏洞的修补,以及安全措施的设置都是非常关键的。而大数据的 4V(大数据量、高速、多样性、真实性)和 1C(复杂性)特征,在公共卫生领域同样存在,新的措施和方法应该被开发出来消除安全威胁与挑战。

采用 Linux 平台能够有更好的安全措施,但更重要的是需要对安全有高度的认识。中国《网络安全法》于 2017 年 6 月 1 日起施行,对网络运行安全提出了要求,对网络信息安全提出了规定,对违反法规的各类行为提出了惩治措施,这也从法律上实现了数据安全。

与此同时,在进行科研数据共享之前,需要执行个人信息的去隐私,保证个人及家庭的数据信息安全。其思路是对每个数据集提供唯一标识,并为数据提供者创立数字认证。对于个人数据,需要移除姓名地址等关键信息,实现个人隐私安全。只有能够全面保护个人隐私,才能更好的实现数据的共享和利用。

9　应用前景

公共卫生大数据分析可以服务于多个不同的方向,为公众卫生水平的提升提供技术指导和数据支持。可预见的应用体现在以下方面。

9.1　大规模流行病预测

通过对大量数据的分析,能够对疾病流行、发展情况进行评估和预测。研究表明,2015 年,全球范围内 11.5% 的死亡原因可归咎于吸烟,而其中 52.2%的死亡发生在中国、印度、美国、俄罗斯等 4 个国家。控烟能产生很好的效果,但也需要全球各个国家的共同努力[19]。心血管疾病中,高血压是重要的因素,而体质指数升高、体力活动减少都是重要诱因。而饮食结构和生活方式改变、快速城市化和工业化则可能是导致中国心血管病剧增的因素[20]。这些结果为制定相关的应对措施提供重要支撑。

2013 出现的 H7N9 流感病毒包含的氨基酸突变,具有哺乳动物的受体结合能力。通过对病毒传播的监测,以及对序列进行的进化分析表明,该病毒可能始于家鸭 H7 病毒,并同 H9N2 病毒株发生重组,进而发生广泛的传播[21]。实际上,病毒传播之前,会有一些线索在各个层次显示出来,例如在小范围内会形成病例增加等现象。应用大数据技术分析活禽交易网络数据,结合 H7N9 毒株的血凝素基因核酸序列构建系统进化树,可推断禽流感疫情在各省及城市间的传播情况,具有较高的应用价值[22]。通过进行大尺度传染疾病的实时监控统计,实现时、空、事件类别的大数据分析好实时监控,能及时提出疫情预报,进而可采取补救措施,分析流行原因,切断传播途径。

9.2　典型疾病的防治

通过公共卫生大数据分析,能够提前预知特定疾病发生、流行的规律,这样就能有效识传播规律,进行有效防治。寨卡病毒被认为是伊蚊传播,引起新生儿小头畸形。通过防止蚊子叮咬、去除蚊虫滋生环境可以进行有效防控。大骨节病是典型的具有地域特点的慢性病,通过对遗传因素、地理环境、饮食结构、基因表达等多层次的研究,识别疾病诱因,可对该病的防治提供科学有益的指导[23-24]。

9.3　健康影响因素的识别

饮食习惯、生活环境会对群体的健康有很大影响,通过大数据分析,可识别影响疾病健康的主要因素。银屑病患者具有较高的代谢病发生率,代谢情况改变同该疾病的病因和治疗、症状密切相关,不良生活习惯,如吸烟、运动减少、肥胖等会增加伴发代谢综合征的概率以及银屑病的病情,导致恶性循环。通过对代谢谱的检测,发现了同疾病相关的差异血清代谢谱,提示在治疗的同时,通过改善饮食结构、生活习惯可减缓疾病的症状[25]。通过大数据分析,不仅能识别疾病的相关因素,还能识别改善措施的效果。AD(老年性痴呆)会随着年龄增长而风险增加。通过对藏族人群 AD 疾病状态的统计和分析,发现藏族特有的宗教行为,包括磕长头、念经、拨念珠等都是 AD 患病的保护因素。这些活动在增加了精细运动和整体运动的强度,使大脑得到了锻炼,加强了神经元之间的联系[26]。每年有 56 万人因不吃水果而死于心血管病,其中 20 万人在 70 岁前死亡。研究人员对45 万中国健康人进行了跟踪随访,发现每天都吃水果的人不但血压和血糖较低,而且得心血管病的人也较少[27]。这些结果使本文有信心对公共卫生大数据进行深入挖掘以识别有效的健康保护因素。

9.4　个人健康保健干预

个人的健康状况影响因素被识别出来后,就可以采取措施,实现更好的健康管理,减少医疗花费,提高生活水平。大规模数据监测有助于制定合理的措施来保护公共健康。1999 年的全国碘营养监测结果发现,儿童尿碘水平为 306 μg/L,处于偏高水平。2000 年中国将生产环节的碘含量出厂不低于 40 mg/kg 下调为平均 35 mg/kg。这样既能向人群提供足够的碘,又把副作用的危险性降至最低水平。缺碘和富碘都会导致甲状腺疾病,沿海地区和内地的膳食中碘摄入量也不同,随着经济社会的不断发展,让民众在知情的前提下进行自由选择,是防治碘缺乏病的有力手段。

代谢是非常关键的生命活动,许多疾病同摄入食品的成分密切相关,糖尿病人不宜多吃甜食是众所周知的,但其他代谢成分对人体的健康并不为人

所知。不同食物的成分和存在的化合物对于慢性病干预和膳食指导也非常关键;当涉及到食品安全问题时候,比如人们摄入被污染或者农药残留超标的食品,将会导致各种急性和慢性疾病。通过大数据的分析,能够及时发现和个人健康相关的影响因素,减少环境因素对身体产生的影响,能及时挖掘到营养素与慢性病之间的关系,及早预防慢性病。

10 遇到的问题和挑战

当前公共卫生大数据的更广泛应用还面临很多问题需要解决,主要体现在以下方面。

1)数据收集。数据的碎片化形式和数据的混杂性特征是数据收集的重要困难。例如,在进行疾病研究时,生存时间是评价治疗效果的重要指标。然而病人的复查信息或身体状态信息往往难以被传递到相应机构,导致随访数据缺失;有些数据需进行提取或格式转换才能用于公共卫生研究,而这时往往缺失统一标准,也难以采用自动化的处理方式,导致数据获取效率低下。智能软件的应用会在数据收集方面提供帮助。

2)隐私保护和数据共享。通常需要合并多个机构的不同数据进行分析,才能获得有效结果,而不同机构的数据格式和内容往往不一致,个人信息通常也不能够被全面获取,同时也难以确定隐私保护的方案。这导致拥有数据的机构难以进行数据分享以及进一步的数据分析。更高程度的信息化有助于这一问题的解决。

3)分析方案的选择和实现。数据之间有千丝万缕的联系,但只有通过合适的分析、统计才能够揭示这些联系。采用 SIR 模型,能够描述一个小区域内某种流行病感染人数的比例。通过结合疾病传播模型,流行病在更大范围的发作情况就能够得到预测[28];利用全球的手术数据,也可以预测哪些地方对何种外科手术有需求,以便制定政策和措施,以满足外科手术治疗需求[29]。其他学科中数据分析方法的引入和借鉴,是解决不同类型公共卫生大数据分析问题的一个重要途径。随着超算和云计算技术的应用,许多占用资源多,耗机时多的方案也能够不断被应用于公共卫生领域。

11 前景分析

公共卫生是居民健康的重要基础和保障。采集到的各种数据资源,连同其他相关数据,形成公共卫生大数据,发挥好这些数据的应用,将产生巨大的社会效益。目前科技的进步正在以全所未有的速度进展,新技术和方法的应用,会不断形成新的成果,覆盖多种公共卫生相关疾病的预警、传播源和传播途径的识别。随着人工智能,机器学习等技术的进步,加上对健康方面知识的积累,以及人们对健康的重视,在提高人们健康水平方面,公共卫生领域大数据的应用将越来越广泛。

在进行公共卫生大数据应用时,需开发科学合理的模型、进行挖据,通过提出假设发现新问题,并利用数据进行推理,获取隐藏在数据中的规律,为最终决策提供支持。但在进行此类研究时候,要充分认识到原始数据的异构性、多样性,数据中干扰因素的存在,以及实现最终应用的复杂性和挑战性。

开展公共卫生大数据的解读分析,需要既懂公共卫生又懂数据分析的“双能”人才。中国人口众多、地域广阔、待解决的问题多样、复杂,急需进行问题的提炼和解决,培养人才队伍相当关键。

可以看到,实现最终目标,还需要多方的努力,包括软硬件,政策环境等的制定。通过协调解决各个方面的问题,公共卫生大数据分析能够发挥更大作用,提升人群健康水平。

12 结 语

公共卫生大数据具有广阔的发展空间,也是解决特定人群健康问题的重要手段之一。采取如下措施,能够促进该方向全面发展。

1)需要形成能够包容多种数据的信息管理平台,提供方便的数据采集和交互。

2)将高性能计算发展成易于广泛使用的形式,形成计算资源的方便使用。

3)数据分析方法作为核心技术,需要能准确提取异构数据中的关键特征。

4)需要培养复合型人才,形成多学科知识的融合。

5)合理、适时的法律法规、政策、标准的制定将对该领域发展有重要影响。以大数据为立足点,多方面的协同将能立体推进公共卫生的健康发展。

参考文献(References)

[1]贺婷, 刘星, 李莹, 等.大数据分析在慢病管理中应用研究进展[J]. 中国公共卫生, 2016, 32(7): 981-984. DOI:10.11847/zgggws2016-32-07-28.
HE Ting, LIU Xing, LI Ying, et al. Application of medical big data in non-communicable chronic diseases management [J]. Chinese Journal of Public Health, 2016, 32(7):

981-984.DOI:10.11847/zgggws2016-32-07-28.

[2] KLONOFF D C. Precision medicine for managing diabetes [J]. Journal of Diabetes Science and Technology, 2015, 9(1):3-7. DOI:10.1177/1932296814563643.

[3] 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 Qiannan, MA Jiaqi. Big data analytics and application in public health[J].China Digital Medicine, 2016, 11(2): 10-12. DOI:10.3969/j.issn.1673-7571.2016.02.003.

[4] GALPERIN M Y, FERNÁNDEZ-SUÁREZ X M, RIGDEN D J. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes [J]. Nucleic Acids Research, 2017, 45(D1): D1-D11. DOI:10.1093/nar/gkw1188.

[5] FABREGAT A, SIDIROPOULOS K, VITERI G, et al. Reactome pathway analysis: a high-performance in-memory approach[J]. BMC Bioinformatics, 2017, 18: 142. DOI: 10.1186/s12859-017-1559-2.

[6] DUMANCAS G G, ADRIANTO I, BELLO G, et al.Current developments in machine learning techniques in biological data mining[J]. Bioinformatics and Biology Insights, 2017, 11: 1177932216687545. DOI: 10.1177/1177932216687545.

[7] MONTAZERI M, MONTAZERI M, MONTAZERI M, et al. Machine learning models in breast cancer survival prediction [J]. Technol Health Care, 2016, 24(1):31-42. DOI:10.3233/THC-151071.

[8] BENSON A R, GLEICH D F, LESKOVEC J. Higher-order organization of complex networks [J]. Science, 2016, 353(6295):163-166. DOI:10.1126/science.aad9029.

[9] GOGOSHIN G, BOERWINKLE E, RODIN A S. New algorithm and software (bnomics) for inferring and visualizing bayesian networks from heterogeneous big biological and genetic data[J]. Journal of Computational Biology, 2017, 24(4):340-356. DOI:10.1089/cmb.2016.0100.

[10] 范婷, 娄岩. 2010-2016 年大数据与健康相关 SCI 论文的聚类分析[J].中国数字医学, 2017, 12(1): 3-5. DOI:10.3969/j.issn.1673-7571.2017.1.001.
FAN Ting, LOU Yan. Cluster analysis on topics of big data and health from 2010 to 2016[J].China Digital Medicine, 2017,12(1):3-5.DOI:10.3969/j.issn.1673-7571.2017.1.001.

[11] TENG Yue, BI Dehua, XIE Guigang, et al. Dynamic forecasting of zika epidemics using google trends[J]. PLoS One, 2017, 12(1):e0165085.DOI:10.1371/journal.pone.0165085.

[12] 任向楠, 丁钢强, 彭茂祥, 等. 大数据与营养健康研究[J]. 营养学报, 2017, 39(1):5-9. DOI:10.3969/j.issn.0512-7955.2017.01.002.
REN Xiangnan, DING Gangqiang, PENG Maoxiang, et al. Big data in the field of nutrition and health[J]. Acta Nutrimenta Sinica, 2017, 39(1): 5-9. DOI: 10.3969/j.issn.0512-7955.2017.01.002.

[13] FUNG I C H, TSE Z, FU K W. Converting big data into public health[J]. Science, 2015, 347(6222):620. DOI: 10.1126/science.347.6222.620-b.

[14] ZHANG Y C, KAGEN A C. Machine learning interface for medical image analysis[J]. Journal of Digital Imaging, 2017, 30(5): 615-621. DOI: 10.1007/s10278-016-9910-0.

[15] PAGE A, OATES S P T, MOHSENIN T. An ultra low power feature extraction and classification system for wearable seizure detection[C]//Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Milan, Italy: IEEE, 2015: 7111-7114. DOI:10.1109/EMBC.2015.7320031.

[16] FERNANDEZ E B, VILLARREAL J, LONARDI S, et al. FHAST: FPGA-based acceleration of bowtie in hardware [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015, 12(5): 973-981. DOI: 10.1109/TCBB.2015.2405333.

[17] 王艺, 任淑霞.医疗大数据可视化研究综述[J]. 计算机科学与探索, 2017, 11(5): 681-699. DOI: 10.3778/j.issn.1673-9418.1609014.
WANG Yi, REN Shuxia. Survey on visualization of medical big data[J]. Journal of Frontiers of Computer Science and Technology, 2017, 11(5): 681-699. DOI: 10.3778/j.issn.1673-9418.1609014.

[18] AFGAN E, BAKER D, Van Den BEEK M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update[J]. Nucleic Acids Research, 2016, 44(W1): W3-W10. DOI:10.1093/nar/gkw343.

[19] COLLABORATORS T. Smoking prevalence and attributable disease burden in 195 countries and territories, 1990-2015: a systematic analysis from the Global Burden of Disease Study 2015[J]. Lancet, 2017, 389(10082): 1885-1906. DOI:10.1016/S0140-6736(17)30819-X.

[20] LI Yanping, WANG Dong, LEY S H, et al. Potential impact of time trend of life-style factors on cardiovascular disease burden in China[J]. Journal of the American College of Cardiology, 2016, 68(8):818-833. DOI: 10.1016/j.jacc.2016.06.011.

[21] LAM T T, WANG Jia, SHEN Yongyi, et al. The genesis and source of the H7N9 influenza viruses causing human infections in China[J]. Nature, 2013, 502(7470): 241-244. DOI:10.1038/nature12515.

[22] 杜鹏程, 于伟文, 陈禹保, 等. 利用系统进化树对 H7N9 大数据预测传播模型的评估[J]. 中国生物工程杂志, 2014, 34(11): 18-23. DOI: 10.13523/j.cb.20141103.

DU Pengcheng, YU Weiwen, CHEN Yubao, et al. Evaluation of the H7N9 transmission model predicted by big data by phylogenetic tree [J]. China Biotechnology, 2014, 34(11):18-23. DOI:10.13523/j.cb.20141103.

[23]郭雄. 大骨节病病因与发病机制的研究进展及其展望[J]. 西安交通大学学报(医学版), 2008, 29(5):481-488.
GUO Xiong. Progression and prospect of etiology and pathogenesis of Kashin-Beck disease[J]. Journal of Xi'an Jiaotong University(Medical Sciences), 2008. 29(5):481-488.

[24]WANG Shuang, GUO Xiong, WU Xiaoming, et al. Genome-wide gene expression analysis suggests an important role of suppressed immunity in pathogenesis of Kashin-Beck disease [J]. PLoS One, 2012, 7(1): e28439. DOI:10.1371/journal.pone.0028439.

[25]姜友贵.基于 GC-MS 寻常型银屑病患者的代谢组学分析[D]. 西安:西安交通大学, 2017.
JIANG Yougui. Metabonomics analysis of patients with psoriasis vulgaris based on GC-MS[D]. Xi'an: Xi'an Jiaotong University, 2017.

[26]尚颖. 青海省 60 岁以上藏族阿尔茨海默病患病率及影响因素研究[D]. 广州: 南方医科大学, 2015.
SHANG Ying. The risk factors of Alzheimer's disease among Tibetan aged 60 years and older in Qinghai Province [D]. Guangzhou: Southern Medical University, 2015.

[27]DU Huaidong, LI Liming, BENNETT D, et al. Fresh fruit consumption and major cardiovascular disease in China [J]. New England Journal of Medicine, 2016, 374(14): 1332-1343. DOI:10.1056/NEJMoa1501451.

[28]PAEZ CHAVEZ J, GOTZ T, SIEGMUND S, et al. An SIR-Dengue transmission model with seasonal effects and impulsive control [J]. Mathematical Biosciences, 2017, 289: 29-39. DOI:10.1016/j.mbs.2017.04.005.

[29]ROSE J, WEISER T G, HIDER P, et al. Estimated need for surgery worldwide based on prevalence of diseases: a modelling strategy for the WHO Global Health Estimate [J]. Lancet Glob Health, 2015, 3(S2): S13-S20. DOI: 10.1016/S2214-109X(15)70087-2.