

以大数据筑牢公共卫生安全网： 应用前景及政策建议

刘 奕

摘 要：大数据是国家提高公共卫生保障能力和应对流行病威胁的关键资源。从可得性和应用价值来看，公共卫生中涉及的大型数据源可以分为五大类，即医学大数据、互联网大数据、地理/气象大数据、基于便携设备的人类行为大数据和零售大数据。将多来源大数据纳入公共卫生体系，有助于改善公共卫生体系信息报送的及时性、完整性以及对新发传染病的监测敏锐性，帮助传染病防控政策精准实施，但面临着数据挖掘和共享、数据集成和处理技术、数据隐私保护以及管理体制和机制不适应等方面的难题。为此，应搭建公共卫生大数据应用协同网络，加大大数据应用专项资金投入，健全公共卫生大数据治理体系，加强专业人才培养。

关键词：公共卫生安全；大数据；公共卫生体系；新冠肺炎疫情

中图分类号：TP311.13 **文献标识码：**A **文章编号：**1003-7543(2020)04-0005-12

随着移动互联网、云计算、物联网和社交媒体等的发展，全球的数据量和数据种类急剧增加，大数据已成为至关重要的经济资产。虽然大数据在打击犯罪、市场营销、金融、气象预测、复杂物理模拟和环境治理等领域已显示出广阔的应用前景，但除了在病毒学和微生物学研究中使用全基因组测序和生物信息之外，整体上看医疗保健领域的大数据应用速度依然较慢。习近平总书记在中央政治局常委会会议研究应对新型冠状病毒肺炎疫情工作时的讲话中明确指出，新型冠状病毒肺炎疫情是对我国治理体系和能力的一次大考，应健全国家应急管理体系，补齐公共卫生短板^[1]。突如其来的新冠肺炎疫情，暴露出我国公共卫生和疾病防控体系建设与经济发展不相适应、重大公共卫生事件初期科学防控预案缺乏、数据共享及转化应用渠道不畅通等

问题^[2]。积极应对将大数据有效纳入传染病监测和调查的挑战，使用人工智能、机器学习等新技术快速而及时地捕捉有关传染性疾病的多来源动态信息，改进公共卫生监测系统和发展预测模型，对于我国在新的历史时期提升应对突发重大公共卫生事件的能力和水平具有重大价值。

一、将新的大数据源纳入公共卫生体系的意义与价值

大数据指的是庞大而复杂的数据集合，具有极大量、多种类、生成速度快、准确性高等特征。围绕信用卡、互联网、手机和社交网络的使用，特别是来自微信、微博、视频和音频网站的非结构化数据在近些年经历了指数级增长。虽然数据挖掘具有探索隐藏模式和规律的巨大潜力，但其本质上是异构的，需要以有组织的形式搜集并予以

基金项目：国家社会科学基金后期资助项目“重塑服务业空间格局：距离、边界与政策的影响”(19FJYB044)；国家自然科学基金面上项目“全球价值链下中国服务业国际竞争力研究：基于贸易增加值的分析”(71573057)。

作者简介：刘奕，中国社会科学院财经战略研究院服务经济研究室主任、副研究员。

整合;加之测量技术、数据存储设备以及链接不同数据集的能力特别是机器学习和人工智能等处理技术应用成本的降低^[3],搜集、组织和分析大数据集以发现有用信息的价值得以极大凸显。公共卫生体系预防疾病、改善健康的能力,很大程度上取决于收集和分析有关疾病和病因的人群水平数据,以及采用循证干预措施的水平。对于传统的传染病监测而言,基于实验室测试和公共卫生机构收集的医学数据是判定的黄金标准,但可能会在时间上有所滞后,而且存在产出成本高、对小区域和部分人群的监控精确性不足等问题。与传统的监控方法相比,源自电子病历、志愿者众包数据以及社交媒体、互联网和手机等其他数字来源生成的大数据,可以提供有关传染病威胁或暴发的更及时、详细的本地化信息,是现有方法的必要补充。

(一)有助于改善公共卫生体系的信息及时性

现有传染病的传统信息积累模型仍然是分级辐射模式的变体,由较小的报告中心向地方一级的公共卫生机构报告,再由其报告至国家级公共卫生机构,累积的信息在总体级别进行汇总、处理、更正后,有关的行动再以自上而下的方式传播到整个系统。从保证流程专业和规范的角度来看,这种方式是有效且经济的。但是,信息传递和处理步骤不可避免地会出现时间延迟,并将导致公共卫生干预的响应效率降低。在应对迅速发展的传染病威胁时,医学评估和循证案例积累的时滞,带来的后果可能是严重的。虽然在国际层面和一些国家层面已经建立起的信息直报系统被认为可以成功识别潜在疾病的暴发并实现疾病暴发时的数据共享,但这种信息积累模型在可预测人群和疫情暴发的情况下才更为有效。在人群大规模流动的情形下,传染病将以意想不到的方式和速度传播,疾病发生和报告之间的时间间隔以及数据空间分辨率的局限性使监控系统的有效性大大降低。相关研究表明,从首次报告到区域或国家开始调查的时间,是疾病暴发严重程

度和总持续时间的重要决定因素^[4]。众包数据以及来自智能手机、社交媒体和互联网等多种不同技术和平台的大数据,可以提供个人水平上的实时信息,虽然无法直接产出准确的病例数,但可以用来增强医疗系统和实验室的监控,有助于从疾病首发到报告监测时间间隔的缩短,从而实现更及时、更有效的传染性疾病管理。研究发现,综合6种基于互联网的传染病大数据分析系统,能够比官方公共卫生监测平均提前10天发现H5N1流感的暴发^[5]。

(二)有助于增进公共卫生体系的信息完整性并提升分析精度

在使用常规传染病报告系统的医疗机构中,可能会发生漏诊、漏报和延误报告的情况,这些信息通常在汇总分析中被掩盖。尽管传统监测系统遗漏的病例在临床上可能不太严重,未被涵盖的偏远地区在统计上也不够显著,但对于了解疾病传播规律、进行预测性建模和控制工作,以及衡量疾病造成的生产力损失可能非常重要。而大数据集的一个突出特点是 $N=全部$,即不需要样本,可以使用整个背景种群,故而不存在抽样偏差问题。而且,智能手机和可穿戴设备提供了有关位置、活动和联系人的大量实时和高分辨率数据,增加了在“难以到达”的人群和地区中获取信息的机会,使得以前所未有的细节和广度深入了解人群行为及其健康风险因素成为可能。比如,借助大数据,决策者可以超越年龄、性别、种族、民族等传统指标来进一步细化患者,并在适当的地理范围内进行汇总,从而反映新发传染病的群体和地理异质性,对不同类型干预措施的潜在效应进行有效评估。

(三)能够提高公共卫生体系对新发传染病的监测敏感性,节约社会成本

传统的传染病监控系统正受到来自新发传染病的挑战,如COVID-19、MERS、SARS、Zika、Ebola等。为了满足发现新疾病特别是发现可能与生物恐怖主义相关疾病的需求,北美、欧洲及

其他地区的许多国家已投入大量资源开发综合症状监测系统。症状监测使用确诊前可辨别的临床特征或症状发作提示的活动作为疾病活动变化的警报,因而可以检测新出现的疾病。除了可以通过大数据平台增加新的信息源外,与传统上用于监测异常活动的统计过程相比,大数据分析和机器学习可以为监测系统提供重要的优势,比如依托贝叶斯模型建立的算法,可以使用当前流行病学背景下的观察数据和疾病先验概率来估计疾病事件实际发生的概率,从而增进公共卫生体系的战略敏捷性。此外,传统的以生物学为重点的监测技术,采取患者访谈和实验室检查、后生成正式报告链条的模式,虽然通常被认为是准确的,但过程成本也是极其高昂的。大数据具有用公式化算法代替和支持人类决策的能力,可以更深入地了解 and 洞察影响健康的因素。与传统的研究方法相比,它能够提高统计的有效性,在公共卫生系统中的应用有望成为大数据领域的最佳实践。与个人生活趋势相匹配时,从电子健康记录中提取的数据不仅能够提升单个患者的护理水平,而且可以发现潜在的健康风险,预测公共健康趋势并在整体上有效降低医疗成本。

(四)能够提升预测模型的精确性,帮助传染病防控政策精准实施

近年来,对个人和人群进行更精确描述分析的“精准公共卫生”,成为公共卫生体系发展的重要方向。精准公共卫生是指通过应用新兴的方法和技术监测人群中疾病、病原体暴露行为和易感性,从而提高预防疾病、促进健康和减少人群健康差异,以及制定改善健康的政策和目标明确的实施方案的能力。作为一种新兴实践,精准公共卫生鼓励使用新数据、新技术和新方法更精细地预测和了解公共卫生风险,并针对更具体和同质的亚人群定制治疗方法。大数据可以通过改善监测和评估效果来促进精准的公共卫生;对于预测模型来说,作为公共卫生体系了解疾病传播情况、

进行疾病管理计划的重要工具,纳入大数据后,公共卫生预测模型的有效性将有较大提升。与市场营销和天气预报等其他领域相比,将大数据应用于传染病动态建模仍处于起步阶段。传染病与气象事件的预测模型之间最大的区别是可用于参数化模型的信息的质量和数量,传染病模型对接触速率参数和组内组间接触的异质性特别敏感。目前大多数公共卫生模型都假设组内进行随机混合并仅关注较大的人口中心之间的混合;而新的大数据分析工具可以为传染病模型提供详细的运动信息,并捕获种群内接触的异质性。为了响应媒体的关注以及由此引起的公众关注变化,这种精确的运动数据还可以测量感染者和易感者之间的不同接触率。通过智能手机和应用程序,还可以收集有关人员、地点和时间的数据,并据此评估其施行效果,调整干预措施。因此,从公共卫生的角度来看,大量数据的收集构成了流行病学研究中不可估量的资源,有助于分析人群的健康需求、评估基于人群的干预措施并制定明智的政策。

二、公共卫生中的大数据:来源及应用前景

将传统传染病监测与大数据集相结合,是公共卫生体系发展的重要方向。一些数据来源如社交媒体、搜索引擎、卫星数据、全球气候、社交媒体、学校/工作出勤、药品销售等本身并未包含很多关于健康的信息,虽然不能替代公共卫生机构和研究人员提供的高质量监控数据,但如果将结构化和非结构化的多个传统与非传统数据源融合在一起,并通过使用一系列有助于从噪声中提取信号的工具,它们所反映出的社区健康状况的图景将比仅从临床数据中获得的更为迅速和可操作。从可得性和应用价值角度来看,公共卫生中涉及的大型数据源大致可以分为五大类,即医学大数据、互联网大数据、地理/气象大数据、基于便携设备的人类行为大数据、零售大数据。

(一) 医学大数据

从来源上看,医学大数据主要包括临床大数据、全基因组测序(WGS)数据和众包大数据。这些大数据在流行病学和分子生物学研究中已得到一定程度的应用,但其在传染病监测和调查中的作用尚待深度挖掘。

1. 临床大数据

在医疗领域,大数据来自医院信息系统、医生的工作、患者体检、药房、治疗记录和医学成像等,产出的主要数据包括电子健康记录(EHR)/电子病历(EMR)、个人健康记录(PHR)和医学图像。EHR 目前已被许多国家采用,2012 年电子健康记录产生了约 500 PB 的数据,2020 年将达到 25 000 PB^[6]。尽管电子记录是公共卫生监测数据的最直接来源,但也有许多因素会限制其利用。比如,这些大数据大多是非结构化的,用于分析来自异构源信息的工具尚在探索之中;高机密性的数据集即使是已取消标识和汇总的数据,隐私问题也会成为访问的重要障碍;而且,EHR 通常不会记录包括环境或行为风险因素等重要的关注变量,在公共卫生监测活动中的用途可能有限,故而大多数电子健康信息系统主要服务于临床,而不是公共卫生。

2. 全基因组测序(WGS)数据

全基因组测序(WGS)数据的分析,是大数据在传染病监控和调查中最广泛认可的应用之一。WGS 可用于从特定暴发调查或疑似来源中吸纳或排除病例,并以此帮助公共卫生系统确定感染控制措施。在某些情况下,还可以使用进化分析方法推断某些病原体的起源和出现,估计潜在的疾病来源并确定最可能的传播链。然而,WGS 的应用仍有其局限性,工具性能的因素有可能导致影响分析的误报,目前无论是 WGS 还是任何单独的技术尚不足以确定传染病暴发及其原因;生物信息学相关研究还对报告数据处理和 WGS 方法的透明性提出了质疑,并提出有必要对用于公共卫生实践的 WGS 方法论和生物

信息学工具进行标准化。

3. 众包大数据

作为一种基于众包数据的主动监测方法,参与式人群监测指的是志愿者通过互联网或电话访谈形式主动提供健康相关信息。参与式监测在志愿者注册时进行背景调查,并对志愿者是否出现某些症状或行为进行连续调查,由此可以对一般人群中疾病或健康事件等分布情况作出估计^[7]。由于相关信息由志愿者主动提供,参与式监测数据较之公共卫生系统被动搜集的可靠性、完整性和针对性都更强。从 2003 年荷兰和比利时最先发起参与式监测系统——大流感调查 Great Influenza Survey 以来,参与式系统已被世界各国广泛用于流感监测,典型的如英国的流感调查 Flusurvey、澳大利亚流感追踪 FluTracking、墨西哥呼吸道疾病监测系统 Reporta 和在美国、加拿大上线的 Flu Near You 等。欧洲流感监测网 Influenzanet 目前已覆盖 10 个国家和地区,有 5 万名注册志愿者,而且其可扩展属性决定了添加额外的参与者不会显著增加成本。对流感网的评估显示,其可以在前哨医师发布之前成功检测出病毒活性的变化^[8]。然而,基于自愿者的监测可能具有样本自我选择偏差、混杂因素等问题,自我报告的形式无法避免故意误报,对于未经医生和实验室测试证实的体征和症状识别也存在一定困难。

(二) 互联网大数据

互联网大数据指的是个人在使用互联网时留下的数字痕迹,其中涉及公共卫生行为的大数据主要包括与公共卫生有关的新闻报道、对健康相关主体信息的网络搜索、微信、微博和网络论坛上有意共享的相关信息等。

1. 汇总新闻报道

事实表明,来自非医学渠道的疫情信息非常重要。目前,世界上使用媒体新闻报道建立的疾病监测系统包括 GPHIN, ProMED-mail 和 HealthMap。监测新兴疾病计划 ProMED-mail 是

最早的基于事件的温和监测系统,信息来源主要是媒体报道、官方报告、网上汇总和当地观察员得到的世界疫情情报。ProMED-mail 目前主要和托管于哈佛大学的 HealthMap 合作,并结合世界卫生组织、GeoSentinel、世界动物卫生组织(OIE)、联合国粮食及农业组织(FAO)、EuroSurveillance、Google 新闻、Wildlife Data Integration Network、百度新闻和 SOSO 信息等多个来源渠道。HealthMap 每天 24 小时监控、汇总和过滤以 9 种语言收集的信息,在一项针对 2012 年 6 个月内 111 例传染病暴发的研究中,HealthMap 在首次正式报告之前的平均 1.26 天($p=0.002$)就发布了预警^[9]。

世界卫生组织早在 20 世纪 90 年代后期就建立了“全球公共健康情报网”(Global Public Health Intelligence Network, GPHIN),总部设在加拿大。GPHIN 使用与各国和地方报纸以及精选新闻通讯链接的新闻聚合器,可以收录和挖掘 9 种不同语言的 30 000 多个全球新闻源,所考察的新闻涵盖健康、体育、旅游和金融等多个领域。GPHIN 每 15 分钟扫描一次,能在不到 1 分钟的时间内完成翻译和数据处理,并结合人工判读识别发现和跟踪公共卫生威胁,以实现基于互联网的实时、早期预警。GPHIN 系统被认为首先预测了 MERS 的暴发,并且通过抗病毒药物销售量增加预测了 SARS 在中国的早期活动^[10]。

除了以上两种代表性的全球监测系统之外,有一些研究还尝试通过对媒体报道的自定义查询预测传染病暴发,但同时也发现了一些应用局限。比如,媒体报道关注耸人听闻故事的特性,决定了监测结果有可能发生偏差,需要和其他医学数据、众包数据和智能手机数据相结合。未来随着人工智能技术的发展,全球监测系统有望通过搜索算法和维护系统的更新,实现功能的不断完善。

2. 搜索引擎

谷歌流感趋势(Google Flu Trends)基于谷歌搜索引擎,每天使用查询日志跟踪特定区域的总

搜索量和输入特定搜索词的频率,据此预测流感的发生时间和规模。谷歌流感趋势的预测效果令人鼓舞,可以做到比美国疾病控制与预防中心(CDC)的 FluView 快 7—10 天^[11];它甚至被证明可以预测肠胃疾病的暴发以及轮状病毒在幼儿中的暴发^[12]。然而,其失守的事件也屡屡提醒人们,使用互联网数据进行传染病风险预测的方法,可能会在少数情况下过度拟合模型。2013 年 2 月,谷歌流感趋势估计的流感相关就诊次数是 CDC 定点诊所和医院报告的两倍^[13]。尽管 Google 流感趋势与实验室确诊病例的相关性一直偏低,但通过与历史监测数据结合,可以显著改善预测精准度。另一个重要的局限是,其在流行病学中解释特定年龄差异的能力有限,当然通过 IP 地址捕获位置等相关信息也是可以考虑的改进方法。

3. 平台大数据

在平台大数据监测传染病方面,已有研究主要关注在线餐厅评论论坛用于估计食源性疾病的风险或监测疾病暴发的可行性问题。其中,Yelp(www.yelp.com)作为最经常使用的饭店评论网站,理所当然成为了主要研究对象。通过回顾顾客在 Yelp 上发表的评论,比如根据对两个或两个以上顾客生病且潜伏期大于 10 小时的“病”“呕吐”“腹泻”“食物中毒”等词的搜索来确定病例,已有研究预测 9 个月内发生的 3 次未报告的与餐厅相关的传染病暴发^[14],从而证明了在线餐馆评论在识别食源性疾病暴发方面的有效性,特别是具有识别传统监测工具无法捕获的小点源疾病暴发的潜力。除了评论文字中的监测线索外,已有研究还通过建立统计模型,考察了评论的数量、长度、平均得分、负面评论的数量以及虚假评论数量,发现评论内容与公共卫生部门卫生检查结果之间的高度相关性^[15]。此外,基于 7 种语言对于维基百科(Wikipedia)数据查询的模型,可以对登革热和流感的发病时间监控前置多达 4 周。目前,美国的一些大城市已经使用 Yelp 或 Twitter 的数据来确定检查的餐厅名录,对于

平台大数据的监测在确定严重违反健康法规的场所方面取得了可喜的成果。可以预期,在我国通过对美团点评、饿了么等平台开展类似的分析,也有望取得较好的监测效果。当然,虚假评论的存在也要求通过电话、发送私人信息等形式对异常评论和评分进行人工复核,或者从其他评论网站或地方卫生和食药监管部门导入数据加以印证。也有研究指出,当以大型连锁餐饮为样本或使用高排名的评论时,模型准确性更高。此外,由于公众缺乏对各种病原体不同潜伏期的专业知识,评论人也有可能将感染源归于错误来源。

4. 社交网络大数据

社交媒体提供了利用公众集体智慧对传染病早期传播进行控制的机会;通过非结构化的文本挖掘,公共卫生管理部门可以获得反映传染病症状的被动监视实时数据。已有研究主要是基于 Twitter 和 Facebook 两个平台展开的,对 Facebook 的研究主要关注慢性疾病的发生风险评估,而 Twitter 则被更多用于监测不同地区的流感、霍乱、大肠杆菌和登革热等传染性疾病的暴发,并根据相关活动解释其原因。除了分析疾病的强度和原因之外,一些研究还使用社交网络分析和空间集群的识别方法,分析了疾病暴发的空间特征^[16]。比如 Broniatowski 等开发了自动将相关推文分类的流感病毒监测系统,并测试了该系统2012—2013年流感季期间多个地理粒度的性能,发现与各级卫生部门监测的流感流行率相比,系统准确度可以达到85%^[17]。还有一些研究通过 Twitter 大数据了解人们的活动模式,用以估计疾病暴发期间人群的接触率^[18],因为与从手机运营商处获得的呼叫记录数据相比,带有地理标签的推文提供的位置数据分辨率更高。此外,一些研究正在尝试将对于 Twitter 建立的模型扩展到 Instagram 使用;还有研究应用社交网络数据评估了流感疫苗的接种情绪,发现消极情绪比积极情绪传播更有效,而且负面情绪较高与疫苗

接种率较低之间存在相关性^[19]。值得注意的是,虽然已有研究证明 Twitter 在公共卫生方面具有广泛的适用性,但使用社交媒体监控传染病的困难也是显而易见的。目前的应用范围局限在学术界,各国卫生部门都未将其大规模用于日常监管实践。推文中使用的语言是动态的,并且可能具有明显的地理异常,需要对同预期不符的结果进行人工检查,不断更新流程的过程分类算法。查找“生病(sick)”之类的简单算法将不起作用,因为“我厌倦了工作”和“我感到不舒服”等推文都不能被解释为与疾病有关,因而需要建立可以将囊括参考原始搜索词但未反映个人疾病的推文内容进行有效分类的高级机器学习算法。而且,社交媒体监视系统的准确性会随着媒体的关注而下降,因为媒体的关注会增加与传染病有关但与实际感染无关的消息,掩盖了真正的流行迹象。此外,社交网络分析的潜在侵入性,也决定了需要在公共利益和保护个人隐私之间进行谨慎权衡。

(三) 地理/气象大数据

在预测传染病方面有价值的地理空间数据包括降水、温度、海拔、土壤类型、植被、永久性和短暂性水体、洪水、土壤湿度和湿地、土地利用,以及人口密度和涉及人口统计学变量的普查数据,这些数据主要来自卫星影像遥感观测和直接野外观测,此外无人机也可能提供新的高分辨率环境数据源。迄今为止,地理大数据在研究中已经用于预测许多媒介传染病,包括裂谷热、埃博拉病毒、登革热、墨累谷脑炎和寨卡病毒。ArcGIS 是 ESRI 为预测分析而开发的一种工具,具有疾病暴发地点查询和基于历史数据的疾病威胁监测等功能,能够通过多次迭代进行假设检验,并生成预测性地图。Attaway 等使用 ArcGIS 分析温度、降水、海拔、植被、人口密度以及其他可从公共来源获得的变量,识别全年适合伊蚊活动的位置及时间,进而识别出可能存在自发性寨卡病毒感染危险的地理区域^[20]。地理大数据还能用来揭

示区域卫生问题,比如杜克大学的研究项目就将数百万患者的记录同地理信息系统数据相结合,通过构建集成的临床数据仓库研究了区域公共卫生的社会决定因素。虽然有关地理大数据的分析过程可以在较短时间内完成,但分析建立在长期数据可用性的基础之上,卫星图像信息在使用前也需要进行大量预处理。

在气候大数据方面,一些研究已经证实,天气是传染病发生的重要影响因素。Anyamba et al. 通过建模计算出撒哈拉以南非洲由于厄尔尼诺现象而存在超过特定阈值的降雨和植被异常,这样的气候变化直接指向裂谷热的暴发;对模型性能的回顾分析表明,模型可以成功预测东非70%和苏丹约50%的裂谷热暴发^[21]。与气候和天气有关的大型数据集可用于传染病传播规律的研究,比如海面温度会影响降水,进而影响地面温度和植被,从而改变传播疾病的媒介及传播条件。通过使用天气大数据、互联网大数据和深度学习,可以更有效地预测传染病。以美国国家航空航天局(NASA)为例,其拥有的数据集包括35年的海表温度和植被模式、37年的降水量和16年的地表温度。此类长期的大型数据集可以监测到气候异常,这些异常虽然本身不甚重要,但可能成为传播媒介的发展条件,对于研究传染病传播规律非常有价值。此外,鉴于空气污染物是触发呼吸系统疾病的已知诱因,各地通过室外传感器搜集的监测空气质量数据,也可以用作呼吸道传染病预测的重要数据源。

(四) 基于便携设备的人类行为大数据

移动网络运营商通常会保留至少3个月的手机通话记录(CDR),通过接入手机反映的基站位置信息,能够获得设备的粗略地理位置;在采取适当措施确保匿名的情况下,来自手机的通话记录数据可以用于位置回溯或者持续追踪。COVID-19疫情发生后,工业和信息化部随即部署了CDR大数据支撑服务疫情防控的相关工作,3家基础电信运营企业基于电子大数据分

析,向用户提供本人14天内到访地查询的服务,有效提升了对流动人员行程查验的效率;另外还针对定点医院、发热门诊、人员聚集区等重点区域的人流变化进行了重点关注,为疫情态势研判和精准防控提供了有力支撑。迄今为止,智能手机上的许多复杂的功能尚未得到充分利用,这些功能用于传染病监视具有巨大潜力。智能手机可以收集、存储和传输GPS坐标,详细的GPS和WiFi数据为捕捉精细的个体运动提供了前所未有的机会;蓝牙传感器可用于跟踪距离,甚至更精细的信号或更细颗粒度的联系网络。而基于加速计、陀螺仪、环境温度和光线等传感器搜集的数据,也可为相关研究提供支持。此外,移动电话可用于解决全球医疗数据不均等问题,特别是对于中低收入国家更为重要,因为在这些国家中,与社会和健康相关的常规数据通常不完整或根本不存在。

公共卫生系统还应重点关注通过可穿戴设备收集的用户生理数据,通常包括个人参数(血压、心率、血糖)、心电图、生命体征、氧气水平、活动信息、传染病、运动和饮食等信息。由于人们越来越关注自己的健康,许多硬件如带有第三方应用程序的智能手机、苹果手表、安卓手表和谷歌眼镜已经在医疗保健领域开发带有传感器的设备,并且在记录个人的日常健康状况以及患者的行为、体征和症状中起着关键作用。此外,人们的运动和饮食数据也可以对评估公共卫生行为作出重要贡献。据不完全统计,仅在Apple iTunes商店中,就有40 000多种医疗保健应用程序可用;到2017年,有超过17亿人下载健康相关应用程序^[22]。将基于便携设备的人类行为大数据纳入公共卫生体系势在必行,但需注意此类大数据存在样本偏差,社会经济地位较低的人、儿童和老年人的代表性可能不足,偏远农村地区的地理位置分辨率较低。在疫情防控期间,需要电信运营商与第三方共享手机通信记录数据,这就存在个人信息保护方面的风险。

(五)零售大数据

在收集零售信息识别传染病暴发方面,药店的处方药和非处方药零售可成为监测数据的重要来源。药品零售大数据在报告时效和地理分辨率方面具有显著优势。以加拿大为例,对于严重的呼吸道疾病监测,加拿大公共卫生局会在症状发作后约10天获得疾病数据,17天后获得实验室数据;但交易完成48小时后就可获得非处方药销售数据。研究表明,季节性抗病毒药物销售与确诊流感病例的发病日期和确诊病例总数密切相关,非处方药销售量与实验室呼吸道病毒病例数和其他呼吸道病毒检出数之间也存在显著关联^[23]。英国的非处方药零售数据目前已用于监测流感活动的时空格局,监测商品涵盖了成人和儿童的感冒药、流感药、咳嗽药、温度计和抗病毒产品(包括洗手液和湿巾纸)。对英国的研究也显示,流感病例与温度计和抗病毒洗手液的销量之间存在显著的正相关,但国家一级的非处方药零售同监测病例的相关性不强,而在更细的空间尺度和年龄组别上,药品零售数据有助于扩大现有的监测范围^[24]。此外,还有研究发现,非处方止泻药和抗恶心药物的销售同诺如病毒、大肠杆菌等活性相关^[25]。除药品监控之外,有关食品销售的零售数据也已被用于调查食源性疾病在人群中的暴发。IBM使用时空分析系统,在考虑产品保质期、可能的消费日期以及产品包含特定病原体可能性的基础上,通过比较零售扫描数据和食品流行病病例的位置,发现了12种最有可能致病的嫌疑食品。当然应用时还需注意,购买行为也许会受到公共卫生信息发布或媒体关注度的影响。

基于大数据构建公共卫生体系的时代已经到来。上述五类大数据远未涵盖可用数据的全部,保险理赔数据、学校出勤数据、工人旷工数据等也已被证明具有较好的监测效果。大数据在传染病监测、增进对疾病的了解、预测未来风险、实行有针对性的干预措施等方面已显示巨大潜力,随

着可用数据规模和种类的增长及数据捕获成本的持续降低,大数据在未来将成为公共卫生体系的必要支柱。

三、将大数据纳入公共卫生体系面临的挑战

虽然上述多来源大数据在传染病监测方面的初步应用较好地拟合了现实,显示巨大的潜力和光明的前景,但将其应用于我国公共卫生管理实践,在数据挖掘和共享、数据集成和处理技术、隐私保护、管理体制和人才储备等方面仍存在不小挑战。

(一)数据挖掘和共享的困难

一方面,数据尚未完全嵌入相关组织的业务流程和组织管理实践中。在许多情况下,患者监测数据并未整合到临床诊断和治疗中,临床数据也未整合到公共卫生服务和传染病监测中,因而需要更多的数据集成。另一方面,对数据的有限访问是实现大数据潜力的主要障碍。公共卫生大数据分散在不同的数据池中,这些数据集之间没有太多联系,由于数据共享机制不完善,医院、科研单位和其他机构之间存在着严重的信息壁垒,在机构之间共享结构化数据效率低下,而在不同组织之间共享非结构化数据则更难以实现。已有的全球疾病监测组织和系统集成度较差,并且生成的数据报告格式不完整、不一致且不兼容。随着数据全球化的加快,公共卫生中的大数据也将面临不同程度的语言、术语和标准化障碍。

(二)数据集成和处理技术的挑战

当前的标准和技术尚不足以满足公共卫生大数据集成应用的要求。数据缺乏统一的标准、一致的描述格式和表示方法,很难实现不同级别的结构化、半结构化和非结构化数据集成;各种数据库使用不同的软件和数据格式,使得数据比较、分析、传输、共享变得非常困难。涉及数据不准确、数据丢失和选择性测量的问题会影响预测建模的结果和决策,模型校准中的缺陷也可能会干扰推断。尽管结合了传统传染病监测方法和大

数据的新型混合模型显示较好的前景,但当前技术手段在处理异构和实时数据方面依然效率不高,很难同时兼顾分布式系统的可用性、一致性和分区容错性,信噪比问题尤其具有挑战性,预测的可靠性尚无法与气候学等领域相比,任何新颖的数据流在投入使用前都必须根据已建立的传染病监测数据和系统进行验证。已有研究表明,将流行弹性行为纳入有望提升模型的解释力,因为从风险感知到勤洗手、戴口罩、改变社交距离和减少出行行为等预防措施的采用,都将显著影响疾病在人群中的传播速度和传播途径。

(三)隐私保护与公共利益间的权衡难题

公共卫生大数据比其他类型的大数据更加敏感,患者数据泄漏可能会带来被孤立、名誉受损、伤害、歧视等严重后果。云计算等大数据技术、医疗保健信息的集中化趋势和一些大型数据库的建立,使得私人医疗数据面临受到攻击的巨大风险,并为第三方出于商业目的在未经授权的情况下滥用患者的健康信息提供了便利。对于数据隐私保护问题,尚无完美的解决方案。即使大数据数据库使用匿名的个人加密数据,仍然存在重新标识用户身份的风险,并且可能通过数据链接技术重新确定个人身份;而完全删除被认为是识别信息的内容,也将限制跨不同来源链接数据的用途。考虑到卫生大数据在道德和伦理等方面的潜在问题,有必要就如何界定其在公共卫生中的伦理用途和不道德用途进行广泛讨论,特别是患者数据隐私风险如何与公共利益之间妥善权衡。

(四)管理体制和人才储备不适应大数据应用的要求

大数据的价值取决于其对公共卫生从业者和政策制定者的效用。目前,公共卫生基础设施尚不具备分析和使用这些异构大数据的能力,跨学科的方法集成不够,同时拥有数据处理技术和医学知识的专业人才非常缺乏。在国家级公共卫生监测系统是监测系统中心的前提下,需要改革管理体制和机制设计,以便将这些非常规来源生成

的大数据转换为可直接操作的信息。地方公共卫生部门存在人力资源和预算限制,妨碍了其访问和有效利用各种大数据资源。为此,应创造适当的制度和政策环境,探索实现共享公共卫生数据的做法。比如,将数字疾病监测组织作为公共卫生系统的正式合作伙伴,以系统的方式将可操作的数据输入监视系统。

四、以大数据筑牢公共卫生安全网的政策建议

大数据是国家提高公共卫生保障能力和应对流行病威胁的关键资源。为充分发挥大数据在筑牢公共卫生安全网中的重要作用,需从加大投入、搭建网络、优化治理和人才培养四方面入手,使大数据更好地服务于公共卫生管理与决策。

(一)搭建公共卫生大数据应用协同网络

近10年间,我国在公共卫生领域建立了各种医疗健康数据采集平台,但缺乏对跨系统多源数据的整合和综合利用,多元化参与机制并未形成。《国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见》从加强公共卫生业务系统建设、建设网络直报系统等方面,为推进公共卫生领域大数据应用指明了方向。下一步应以出台专项文件为指向,由国家相关部门牵头,整合社会公共信息资源,就大数据在公共卫生领域的应用展开多机构跨学科联合攻关。应在法律上明确限定数据共享范围、方式和内容的前提下,搭建大规模协作网络,创建和动员开放数据、开放学科、开放源代码社区和开放协作平台,协调分析和处理公共卫生大数据的能力,确保在共享知识和数据的环境下形成真正的“研究社区”;定期发布公共卫生领域大数据应用的最佳实践,推动相关研究从数量、深度到应用广度的飞跃。

(二)加大公共卫生大数据应用专项资金投入

近年来,各国均通过优先项目的形式,努力探索公共卫生领域应用大数据的潜力。欧盟委员会于2016年以建立全面的国家公共卫生数据仓库的名义,收集医生记录、住院、药物处方、实验

室和放射学分析及其他广泛的数据类型;美国国立卫生研究院大数据知识(BD2K)计划则使生物医学科学家能够更充分地利用研究界正在生成的大数据。我国应设立专项资金和项目,支持多来源的国家和区域公共卫生数据仓库建设,加快公共卫生大数据的集成和共享。应面向全国各行业征集公共卫生大数据创新产品和服务,重点资助利用大数据开展重点人群防控、传染病信息监测、传染病舆情监测、社区疫情排查、多源数据传染病跟踪分析的智能化公共卫生监控和解决方案,以及为抗体研发、医疗救护、疫情防控、防疫物资调动等提供算力支撑的云计算公共服务平台等。

(三)健全公共卫生大数据治理体系

首先,应要求利益相关者协作并调整其系统的设计和性能,重点就相关技术标准达成一致,包括建立统一的疾病诊断编码、临床医学术语、检查检验规范、药品应用编码、信息数据接口和传输协议等,促进健康医疗大数据产品、服务流程标准化,以实现公共卫生信息和创新技术的最大创新潜力。其次,在事关公共安全的特殊前提下,个人信息的公共属性虽极大凸显,但仍应明确不需个人同意而进行个人信息收集和利用的“例外规则”,并以法律形式规定适用的法定情形。要专门制定应对疫情等特殊情形下采集和使用用户数据的隐私保护标准,将公民隐私权的让渡严格限于防控疫情的目的和范围,并保证政府、研究机构、救援机构、媒体和个人对数据的采集、处理、分析和对分析结果的共享在法律规定的的基础上进行。最后,从保护个人健康信息机密性、隐私性和安全性出发,明确个人医疗保健信息方面的法律适用,采取有力的技术和组织措施防止未经授权的访问和使用。在这方面,各国主要采取两种模式:一种是政府根据基本隐私法专门制定法律法规,如美国的HIPAA,澳大利亚的《健康记录和信息隐私法》等;另一种是将个人医疗保健信息作为个人信息或敏感信息的一部分,通

过法律来保护个人信息或敏感信息,如英国的《数据保护法》以及加拿大的《个人信息保护和电子文件法》等。而具体到此类信息在公共卫生领域的应用,需制定专门的法律法规平衡患者隐私保护同出于公共卫生研究和监测目的的数据共享。

(四)加强专业人才培养

由于大多数参与公共卫生决策制定和实施的专业人员都不熟悉大数据科学,因而为其提供相关领域的职业培训至关重要。在公共卫生管理部门中,应考虑招募顶尖技术专家担当首席公共卫生数据科学家,为构建大数据基础上的公共卫生体系提供人才和技术支持。要对医疗保健从业人员开展数据标准化专项培训,倡导和采用标准代码和标识符。鼓励高校开设公共卫生与大数据技术应用专业,大力培养既熟悉公共卫生业务又擅长大数据工程技术的高水平复合型人才;改进卫生信息学相关专业设置,加大对大数据科学课程的学习力度。对公共卫生相关专业的学生加强数据科学技术的培训,引入大数据分析和相关课程;鼓励引进企业师资,建设校企共建实训基地。 **Reform**

参考文献

- [1] 习近平.在中央政治局常委会会议研究应对新型冠状病毒肺炎疫情工作时的讲话[EB/OL].(2020-02-04)[2020-03-15].http://www.qstheory.cn/dukan/qs/2020-02/15/c_1125572832.htm.
- [2] 丁蕾,蔡伟,丁健青,等.新型冠状病毒感染疫情下的思考[J].中国科学:生命科学,2020(3):247-257.
- [3] 渠慎宁,杨丹辉.突发公共卫生事件的智能化应对:理论溯源与趋向研判[J].改革,2020(3):14-21.
- [4] WALDER C, OSGOOD N, SEITZINGER P. Big data for infectious diseases surveillance

- and the potential contribution to the investigation foodborne disease in Canada: An overview and discussion paper[R]. Report for the National Collaborating Centre for Infectious Diseases, May 2017.
- [5]BARBOZA P, VAILLANT L, MAWUDEKU A, et al. Evaluation of epidemic intelligence system integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events[J]. PLoS One, 2013, 8(3): e57252.
- [6]FELDMAN B, MARTIN E, SKOTNES T. Big data in healthcare: Hype and hope[J]. Dr. Bonnie, 2013(1): 122–125.
- [7]赵永谦, 马文军. 基于互联网信息的公共卫生监测进展[J]. 中华流行病学杂志, 2017(2): 272–276.
- [8]GUERRISI C, TURBELIN C, BLANCHON T, et al. Participatory syndromic surveillance of influenza in Europe[J]. Journal of Infectious Diseases, 2016, 214(4): S386–S392.
- [9]BAHK C Y, SCALES D A, MEKARU S R, et al. Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting[J]. BMC Infectious Diseases, 2015, 15(1): 135.
- [10]DION M, ABDELMALIK P, MAWUDEKU A. Big data and the global public health intelligence network(GPHIN)[J]. Can Commun Dis Rep, 2015, 41(9): 209–214.
- [11]CAMEIRO H A, MYLONAKIS E. Google trends: A web-based tool for real-time surveillance of disease outbreaks[J]. Clinical Infectious Diseases, 2009, 49(10): 1557–1564.
- [12]RISHI D, LOPMAN B, SHIMSHONI Y, et al. Use of internet search data to monitor impact of rotavirus vaccination in the United States[J]. Clinical Infectious Diseases, 2012, 54(9): 115–118.
- [13]LAZER D, KENNEDY R, KING G, et al. The parable of Google Flu: Traps in big data analysis[J]. Science, 2014, 343(6176): 1203–1205.
- [14]HARRISON C, JORDER M, STERN H, et al. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York city, 2012–2013[J]. Morbidity and Mortality Weekly Report, 2014, 63(20): 441–445.
- [15]KANG J, KUZNETSOVA P, LUCA M. Where Not to eat? Improving public policy by predicting hygiene inspections using online reviews[R]. Conference: EMNLP 2013, Seattle, USA.
- [16]THAPEN N, SIMMIE D, HANKIN C, et al. Defender: Detecting and forecasting epidemics using novel data: analytics for enhanced response[J]. PLoS ONE, 2015, 11(5): e0155417.
- [17]BRONIATOWSKI D, MICHAEL P, DREDZE M. National and local influenza surveillance through twitter: An analysis of the 2012–2013 influenza epidemic[J]. PLoS One, 2013, 8(12): e83672.
- [18]JURDAK R, ZHAO K, LIU J, et al. Understanding human mobility from Twitter[J]. PLoS ONE, 2015, 10(7): e0131469.
- [19]KASS-HOUT T, ALHINNAWI H. Social media in public health[J]. British Medical Bulletin, 2013, 108(1): 5–24.
- [20]ATTAWAY, D, WATERS N, GERAGHTY E, et al. Zika virus: Endemic and epidemic ranges of aedes mosquito transmission[J].

- Journal of Infection & Public Health, 2017, 10(1): 120–123.
- [21] ANYAMBA A, CHRETIEN J, SMALL J, et al. Prediction of a rift valley fever outbreak [J]. Proceedings of the National Academy of Sciences, 2009, 106(3): 955–959.
- [22] AITKEN M, GAUNTLETT C. Patient apps for improved healthcare: from novelty to mainstream[J/OL]. IMS Institute for Healthcare Informatics. (2013–10–01) [2020–03–15]. <https://www.mendeley.com/catalogue/patient-apps-improved-healthcare-novelty-mainstream/>.
- [23] MUCHAAL P, PARKER S, MEGANATH K, et al. Evaluation of a national pharmacy-based syndromic surveillance system [J]. Relevé Des Maladies Transmissibles Au Canada, 2015, 41(9): 203–206.
- [24] TODD S, DIGGLE P, WHITE P, et al. The spatiotemporal association of non-prescription retail sales with cases during the 2009 influenza pandemic in Great Britain[J]. BMJ Open, 2014, 4(4): e004869.
- [25] EDGE V, POLLARIF, LAI K, et al. Syndromic surveillance of Norovirus using over-the-counter sales of medications related to gastrointestinal illness [J]. The Canadian Journal of Infectious Diseases & Medical Microbiology, 2006, 17(4): 235–241.

Building Public Health Safety Net with Big Data: Application Prospective and Policy Suggestions

LIU Yi

Abstract: Big data is the key preparation currency for countries to improve public health security capacity and respond to the threat of epidemics. From the perspective of availability and application value, big data sources involved in public health can be divided into five categories: medical big data, Internet big data, geographic/meteorological big data, human behavior big data based on smart phones and retail big data. The integration of multi-source big data into the public health system will help to improve the timeliness and integrity of the information submission of the public health system and the sensitivity of monitoring new infectious diseases, and help the precise implementation of infectious disease prevention and control policies. However, it is facing the difficulties of data mining and sharing, the challenges of data processing methods, data privacy protection problems, and the inadaptability of the management system and mechanism. Therefore, in policy, we should build a collaborative network of public health big data application, increase the special fund and projects on big data application, improve the governance of big data and strengthen the training of professional talents.

Key words: public health and safety; big data; public health system; COVID-19

(责任编辑:罗重谱)