

# 疾病预测中的大数据

■本报记者 胡珉琦

医疗与健康应该放在同样重要的位置。事实上,疾病防治关口前移,能大大节省对医疗资源的消耗。数据分析有助于对老百姓进行健康教育,提升疾病预防效果。尽管在医疗领域,国内还处于数据原始积累阶段,但尝试着用数据挖掘开发项目和产品来占领先机也未尝不可。

2009年,谷歌曾因成功预测冬季流感的传播,从而刮起了一阵大数据变革公共卫生的浪潮。近来,谷歌公布消息,Google X最新孵化了一个名为Baseline的医疗健康项目,用大数据来预防癌症。该项目被认为可能是谷歌有史以来最具野心,也是最困难的项目。

## 收集健康数据预防疾病

根据《华尔街日报》的报道,谷歌认为,多数已经发现的生物标志都与晚期疾病有关,因为这种研究普遍集中于病人。但是,利用现有数据尽早判断疾病的效果不佳。而Baseline旨在帮助研究人员更早地发现心脏病和癌症的各种迹象,进而推广预防措施,并不仅仅把精力放在治疗上。

## 微信

### 治理科普乱象须多管齐下

“手机挂胸前导致心脏病和内分泌紊乱,放裤袋会杀精,常用手机的人患上眼癌的机会比其他人多出三倍”“微波炉加热后的水洗花花死,泡豆子不发芽,对人体有害”“口服风油精可以防中暑、根治狐臭、治便秘”……刚刚过去的7月,“科学”流言榜上又添许多五花八门的“科普贴士”,这些一本正经的“关心”借科普之名,在微信朋友圈、微博等平台上畅行无阻。

其实又何止一两个伪科学的帖子,一些在科学上存在严重漏洞的“科普图书”通过正规出版社出版后大行其道已不是什么新鲜事,尤以涉猎养生保健、外星探秘居多,其为吸引眼球博取利益而无视科学的行径,如同在光天化日之下实施诈骗。

中科院院士周忠和在与笔者交流时认为,科普既需要国家有关部门来做,也离不开民间的参与,但民间科普的乱象丛生,也需要规范。

科普乱象层出不穷,究其原因在于,在我国当前的制度和氛围下,“生产”伪科学的成本太低,而揭发检举的代价太高。除此之外,相应的赏罚机制不健全也是乱象迭起的重要原因。

套用传播学者罗斯的谣言公式“谣言=(事件的重要性×事件的模糊性)÷公众批判能力”,可以知道,只要一件事足够重要(如关乎健康、生死),再配以似是而非的言辞,在我国现有的公众批判能力下,哪怕很拙劣的伪科学都能像病毒一样迅速蔓延。而反观当前的制度环境和文化氛围,对伪科学的监管、打假的力度都不够。譬如,出版社出版一本书,其科学性如何,鲜有专业机构或行业协会的鉴定或评判;一条流传甚广的谣言,基本没人去追究它的源头,更遑论追责了。

其实这种文化现象也存在于严肃的学术圈。一项发表在Nature、Science上的成果一经公布,马上被冠以“很牛”的光环,往往是赞美之词一浪高过一浪,国内的媒体甚至相关科研领域都难以比较客观的评判,更不必说是公开的质疑声或者泼冷水。这其实说明了国人还不习惯质疑的态度、整体文化氛围还产生不了批判的土壤。

我国的科普教育正所处的这种尴尬阶段并非不可理解,毕竟现代科学进入中国才不过百余年,何况这期间还有战争和各式各样的“运动”。然而时至今日,国民科学素养正待全面提高,如不加大力度规范科普市场,科技发展的脚步也势难加速。

书评和文评作为一种文化现象,对图书的可读性以及科学性具有一定的监管功能。然而发表书评等多为民间自发行为,并不能完成对读物彻底地客观评价,更不能覆盖所有读物。鉴于此,政府、科研机构和行业协会要真正履行其应有的职责,组织相应的监管机构和队伍,将科学性纳入一般读物的评价体系,作为其准入市场的一个指标。

此外,在相应的制度建设上,政府应采取积极措施,降低“打假”的风险和代价,鼓励专业人士抵制伪科学。尽管这涉及文化土壤的改良,但如果制度健全、落实得力,在若干年内营造出踊跃打假的风气也并非不可期。

目前一些民间机构参与科普行动,建设科普场馆,这顺应了市场需求。但民间机构组织的青少年科普活动要讲究科学专业,更要寓教于乐。

任何市场都不是彻底的自由市场,既要按市场规律办事,也需要政府部门和社会各界(包括消费者)的监督,科普同样如此。要让市场规律更好地发挥作用,就需要相关部门积极主动地干预和管控,激浊扬清,让好的科普机构、活动、读物享有市场红利,粗制滥造者,就让它被市场淘汰。

象,进而推广预防措施,并不仅仅把精力放在治疗上。

为了完成这一项目,谷歌将从175人那里匿名搜集基因和分子信息,之后还会再搜集数千人的相关数据。目的在于全面描绘健康人的身体究竟应该是何模样。

收集的内容包括尿液、血液、唾液和眼泪等体液,数据将包括参与者的整个基因组、父母的遗传史信息,以及他们如何代谢食物、营养和药物,在压力之下他们心跳速度,化学反应如何改变他们的基因行为。这项研究还将创建一个参与者组织样本的知识库。

之后,谷歌便会利用其庞大的计算能力来寻找这些信息中隐藏的“生物标记”,从而帮助医疗研究人员提前发现疾病。

例如,该研究可能会发现一些能够帮助人们分解高脂肪食物的生物标记。拥有这些生物标记的人,可以将患上高胆固醇和心脏病的时间延后,而没有这类生物标记的人则可能更早患上心脏病。一旦Baseline发现了这一标记后,研究人员便可通过检查了解哪些人缺乏这类标记,并帮助他们纠正习惯,或者开发出新的治疗方法,帮助其更好地分解高脂肪食物。

## 如何定义“健康”

在医疗领域,主流的科学研究旨在精确定位与某种疾病相对应的生物标记,从而作为疾病预测、诊断的依据。在中国科学院北京基因组研究所副研究员聂凌虎看来,Baseline放弃了传统医学精确“打击”的思路,而是利用海量存储和计算能力对人体内所有的基因和分子信息进行轰炸式搜索。依据数据统计,得到健康人群的人体结构数据,进而找出比个体所有分子层面的异常。

他告诉《中国科学报》记者,与已知疾病相关的生物标记只是异常中的一小部分,如果将关注的异常参数范围扩大,或许就可以找到病灶出现前的蛛丝马迹。

“过去,医学往往关注的是疾病患者为什么患病,而谷歌想做的是,搞清楚健康人群究竟是如何保持健康的,并想方设法让那些非健康状态的人群(不一定是疾病患者)调整到健康状态。”

聂凌虎认为,谷歌有全球最大的计算机和数据中心,可以迅速提供搜索结果,运行大量数据的服务,庞大的数据网络有能力用于存储和

计算医疗信息,可以为该项目运行提供支持。

目前,Baseline项目研究还处于探索阶段,但它首先面临如何定义“健康”的问题。

“一个唐氏综合征患儿的母亲可能认为她的孩子是健康的,一个人致残后可能会适应生活并重新定义健康。Baseline定义健康也会面临同样的问题。目前与Baseline相关的报道尚没有足够信息显示谷歌是如何定义健康的。”芝加哥大学卫生促进研究中心博士后、卫生经济学家陈苗在接受《中国科学报》记者采访时表示。

“另一个问题是数据的代表性和不同人群的差异。”陈苗说道,“几千个参试个体难以代表所有人群,特别是不同人群对环境、气候有着不同的生理上的适应反应。比如藏区原住民携带EPAS1基因的一种变异从而能够应对相对低氧的环境;镰刀型红细胞疾病的出现在某种程度上是人群对疟疾的应对。”

此外,他还表示,基因、分子层面的缺陷只是疾病发生的一部分,行为和环境因素对健康的影响也相当重要。当下人体各项参数正常,并不代表未来一定不会得病,反之亦然。因此,当前,疾病预测、预防还无法做到准确可靠。

## “谷歌流感趋势”缺陷

事实上,谷歌在医疗大数据预测方面的尝试早在2009年就开始了,“谷歌流感趋势”项目准确预测了2009年的流感流行,反响巨大。但人们有所不知的是,“谷歌流感趋势”之后的表现并不尽如人意。在2011年至2013年间,该项目多数高估了类流感发病率。陈苗曾撰文分析了其中的原因。

该项目背后的原理似乎很简单:如果某地流感开始流行,那么相关疾病的谷歌搜索就会增多。但现实远比原理复杂。

陈苗解释,谷歌搜索的结果和使用人群“自我估计”的流感发病率相关。但在现实中,普通感冒在不同个体身上的严重程度上会有差别,有些患者因为担心,便搜索了相关信息。这些搜索构成了类流感“自我估计”中的“噪声”。如果只用“谷歌流感趋势”预测流感发病率就会存在问题。

他认为“谷歌流感趋势”项目失误的另外一个原因是在数据分析方面存在欠缺。哈佛大学Gary King和搭档的研究就指出了谷歌项目在分析上的不足。首先,这个项目没有尽可能利用传

统的统计分析手段来剔除系统误差,比如残差的自相关性和季节性;其次,“谷歌流感趋势”忽略了其他数据的作用。如果结合美国疾控中心的数据,谷歌的预测其实有可能被大大改善。

## 隐私安全风险

尽管这些项目还存在缺陷,陈苗认为这并不足以否定大数据对国民健康的潜在贡献。美国卡罗来那医疗系统(Carolinas Healthcare System)已经尝试用信用卡数据辅助其他数据预测急诊使用概率,以及用临床数据预测再入院率。“只是,任何新生事物都有一个发展的过程。大数据预测功能在医疗上的价值想要达到人们理想中的要求恐怕还需假以时日。”

但也有观点指出,目前国内医疗问题还没有解决,谈健康显得有点“奢侈”,还是应该把焦点放在医疗上。

聂凌虎并不这么认为。“医疗与健康应该放在同样重要的位置。事实上,疾病防治关口前移,能大大节省对医疗资源的消耗。”聂凌虎说,“数据分析有助于对老百姓进行健康教育,提升疾病预防效果。”

陈苗也表示,虽然在医疗领域国内还处于数据原始积累阶段,但尝试着用数据挖掘开发项目和产品来占领先机也未尝不可,当然,具体项目需要具体评估。不仅如此,他还提到,政府和行业组织在标准化方面应该先行一步,吸收国外的经验和教训,以免将来面临多个数据标准互不兼容的问题。

此外,陈苗强调,在医疗大数据领域需要格外注意隐私及附带的的海关安全问题。

谷歌掌握上千人的身体结构,小到细胞内部的分子,未来,这些数据对于保险公司而言非常宝贵,它们可以利用这些数据降低风险。

对此,谷歌表示,Baseline采用的是匿名方式,搜集的数据也仅限于医疗目的。这些数据不会与保险公司分享。Baseline还将由机构审查委员会进行监测,监督所有涉及人类的医学研究行为。一旦开启全面研究,委员会将由杜克大学和斯坦福大学医学院组成,监控如何使用这些信息。

“即便如谷歌所言,项目参与人群的社会安全号和其他可能确定个人身份的信息不会被传送到谷歌,Baseline测量的大量的生物标记也使确定个人身份成为可能。”在陈苗看来,为保证个人身份安全,对数据安全也将提出更高要求。

## 极客酷品

### 人造绿叶

在太空零重力环境下,植物很难生长,不过为了利用绿叶生产氧气,可以将提取出来的叶绿素重新“包装”到丝蛋白内,制成人工绿叶,在有水有光的情况下就会产生氧气。它们可以附着在大楼外侧,形成人工绿色植被。



### 无线充电的电动自行车

这辆无线充电的自行车配置线圈与电磁转换器,当骑着它通过铺设发射电磁的线圈道路时,磁场会被激活,进而产生电流被自行车所使用与储存,因此可以大幅减少电动自行车的体积与重量。



### 家用热力发电壶

热力发电水壶可将热水产生的热能转换为电能,在家里停电或野外露营时为手机、平板等电子设备提供持续的电能。它的电池单次可存储10瓦电力,应急之用还是很合适的。



### 便捷过滤嘴

这款安装在标准PET塑料瓶口的便捷过滤器,其制造简单,成本低廉,使用方便,对杂质、细菌的过滤效果很明显,可在外出时随身携带,以备不时之需。



### 翻折花盆

这款翻折花盆可像翻袖子口一样从低矮的小花盆变成宽口的大号花盆,如此一来便不必担心因重新栽培而弄伤植物,也减少了额外花销。



### 咖啡机闹钟

与内置提示器的自动式咖啡机不同,它可以释放缭绕的香气与舒缓的声音来唤醒主人。整套设备采用了感应加热,在水煮沸之后左边的不锈钢球便会互相碰撞发出声音,与此同时,右侧的滴滤系统也会准备好一杯香气宜人的咖啡。



(朱香)

## 看图



### 透明青蛙初成长

日本三重县鸟羽市水族馆最初获赠的30只透明蝌蚪中,有25只已经长成了青蛙。据日本《产经新闻》报道,这些青蛙的体长约为1.5厘米,由于隐性遗传和突变导致色素异常。它们在蝌蚪时期身体呈透明,长大后皮肤开始泛白,但仍可以看见微黑的内脏。这种颜色的青蛙非常少见。图片来源:百度图片