第5章

大数据促进医疗与健康

【导读案例】

大数据变革公共卫生

2009年出现了一种新的流感病毒甲型H1N1,这种流感结合了导致禽流感和猪流感病毒的特点,在短短几周之内迅速传播开来(图5-1)。全球的公共卫生机构都担心一场致命的流行病即将来袭。有的评论家甚至警告说,可能会爆发大规模流感,类似1918年在西班牙爆发的影响了5亿人口并夺走了数千万人性命的大规模流感。更糟糕的是,我们还没有研发出对抗这种新型流感病毒的疫苗。公共卫生专家能做的只是减慢它的传播速度。但要做到这一点,他们必须先知道这种流感出现在哪里。

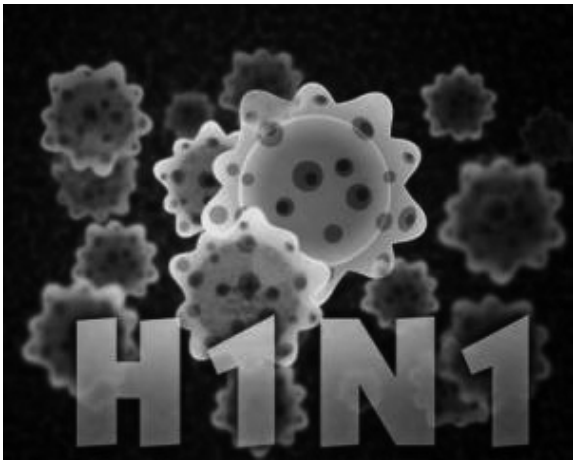

图5-1　甲型H1N1流感疫情全球流行示意图

美国,和所有其他国家一样,都要求医生在发现新型流感病例时告知疾病控制与预防中心。但由于人们可能患病多日,实在受不了了才会去医院,而这个信息传达回疾控中心也需要时间,因此,通告新流感病例时往往会有一两周的延迟,而且,疾控中心每周只进行一次数据汇总。然而,对于一种飞速传播的疾病,信息滞后两周的后果将是致命的。这种滞后导致公共卫生机构在疫情爆发的关键时期反而无所适从。

在甲型H1N1流感爆发的几周前,互联网巨头谷歌公司的工程师们在《自然》杂志上

发表了一篇引人注目的论文。它令公共卫生官员们和计算机科学家们感到震惊。文中解释了谷歌为什么能够预测冬季流感的传播：不仅是全美范围的传播，而且可以具体到特定的地区和州。谷歌通过观察人们在网上的搜索记录来完成这个预测，而这种方法以前一直是被忽略的。谷歌保存了多年来所有的搜索记录，而且每天都会收到来自全球超过30亿条的搜索指令，如此庞大的数据资源足以支撑和帮助它完成这项工作。

谷歌公司把5000万条美国人最频繁检索的词条和美国疾控中心在2003—2008年间季节性流感传播时期的数据进行了比较。他们希望通过分析人们的搜索记录来判断这些人是否患上了流感，其他公司也曾试图确定这些相关的词条，但是他们缺乏像谷歌公司一样庞大的数据资源、处理能力和统计技术。

虽然谷歌公司的员工猜测，特定的检索词条是为了在网络上得到关于流感的信息，如“哪些是治疗咳嗽和发热的药物”，但是找出这些词条并不是重点，他们也不知道哪些词条更重要。更关键的是，他们建立的系统并不依赖于这样的语义理解。他们设立的这个系统唯一关注的就是特定检索词条的使用频率与流感在时间和空间上的传播之间的联系。谷歌公司为了测试这些检索词条，总共处理了4.5亿个不同的数学模型。在将得出的预测与2007年、2008年美国疾控中心记录的实际流感病例进行对比后，谷歌公司发现，他们的软件发现了45条检索词条的组合，将它们用于一个特定的数学模型后，他们的预测与官方数据的相关性高达97%。和疾控中心一样，他们也能判断出流感是从哪里传播出来的，而且判断非常及时，不会像疾控中心一样要在流感爆发一两周之后才可以做到。

所以，2009年甲型H1N1流感爆发的时候，与习惯性滞后的官方数据相比，谷歌成为了一个更有效、更及时的指示标。公共卫生机构的官员获得了非常有价值的数据信息。惊人的是，谷歌公司的方法甚至不需要分发口腔试纸和联系医生——它是建立在大数据的基础之上的。这是当今社会所独有的一种新型能力；以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务或深刻的洞见。基于这样的技术理念和数据储备，下一次流感来袭的时候，世界将会拥有一种更好的预测工具，以预防流感的传播。

阅读上文，请思考、分析并简单记录：

(1) 谷歌预测流感主要采用的是什么方法？

答：_____

(2) 谷歌预测流感爆发的方法与传统的医学手段有什么不同？

答：_____

(3) 在现代医学的发展中,你认为大数据还会有哪些用武之地?

答:

(4) 请简单描述你所知道的上一周内发生的国际、国内或者身边的大事。

答:

5.1 大数据与循证医学

大数据与循证医学

循证医学(图 5-2),意为“遵循证据的医学”,又称实证医学,其核心思想是医疗决策(即病人的处理、治疗指南和医疗政策的制定等)应在现有的最好的临床研究依据基础上做出,同时也重视结合个人的临床经验。

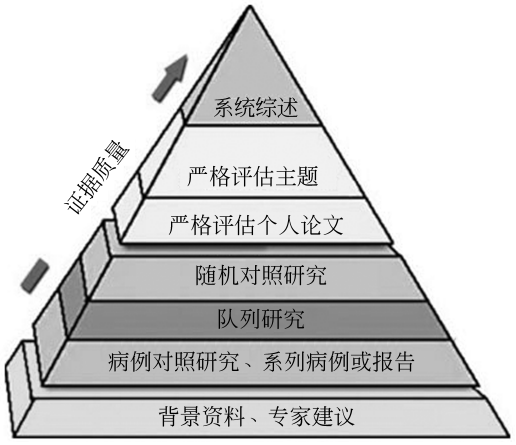

图 5-2 循证医学金字塔

第一位循证医学的创始人科克伦(1909—1988),是英国的内科医生和流行病学家,他1972 年在牛津大学提出了循证医学思想。第二位循证医学的创始人费恩斯坦(1925—),是美国耶鲁大学的内科学与流行病学教授,他是现代临床流行病学的开山鼻祖之一。第三位循证医学的创始人萨科特(1934—)也是美国人,他曾经以肾脏病和高血压为研究课题,先在实验室中进行研究,后来又进行临床研究,最后转向临床流行病学的研究。

就实质而言,循证医学的方法与内容来源于临床流行病学。费恩斯坦在美国的《临床药理学与治疗学》杂志上,以“临床生物统计学”为题,从 1970 年到 1981 年的 11 年间,共发表了 57 篇连载论文,他的论文将数理统计学与逻辑学导入临床流行病学,系统地构建

了临床流行病学的体系,被认为富含极其敏锐的洞察能力,因此为医学界所推崇。

传统医学以个人经验、经验医学为主,即根据非实验性的临床经验、临床资料和对疾病基础知识的理解来诊治病人(图 5-3)。在传统医学下,医生根据自己的实践经验、资深医师的指导、教科书和医学期刊上零散的研究报告为依据来治疗病人。其结果是:一些真正有效的疗法因不为公众所了解而长期未被临床采用;一些实践无效甚至有害的疗法因从理论上推断可能有效而长期广泛使用。

图 5-3　传统医学是以经验医学为主

循证医学不同于传统医学。循证医学并非要取代临床技能、临床经验、临床资料和医学专业知识,它只是强调任何医疗决策应建立在最佳科学研究证据基础上。循证医学实践既重视个人临床经验,又强调采用现有的、最好的研究证据,两者缺一不可(图 5-4)。

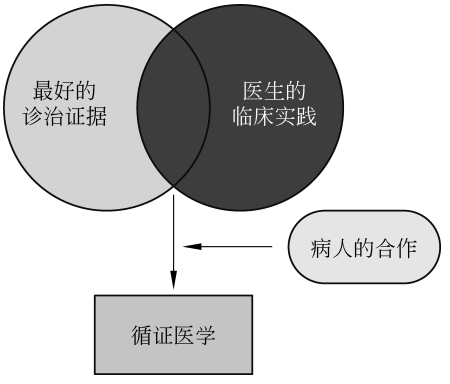

图 5-4　循证医学重视个人临床经验,也强调研究证据

1992 年,来自安大略麦克马斯特大学的两名内科医生戈登·盖伊特和大卫·萨基特发表了呼吁使用“循证医学”的宣言。他们的核心思想很简单。医学治疗应该基于最好的证据,而且如果有统计数据,最好的证据应来自对统计数据的研究。但是,盖伊特和萨基特并非主张医生要完全受制于统计分析,他们只是希望统计数据在医疗诊断中起到更大的作用。

医生应该特别重视统计数据的这种观点,直到今天仍颇受争议。从广义上来说,努力推广循证医学,就是在努力推广大数据分析,事关统计分析对实际决策的影响。对于循证医学的争论,在很大程度上是关于统计学是否应该影响实际治疗决策的争论。当然,其中很多研究仍在利用随机试验的威力,只不过现在的风险大得多。由于循证医学运动的成功,一些医生在把数据分析结果与医疗诊断相结合方面已经加快了步伐。互联网在信息追溯方面的进步已经促进了一项影响深远的技术的发展,而且利用数据做出决策的过程也达到了前所未有的速度。

5.2 大数据带来的医疗新突破

根据美国疾病控制中心的研究，心脏病是美国的第一大致命杀手，每年250万的死亡人数中，约有60万人死于心脏病，而癌症紧随其后（在中国，癌症是第一致命杀手，心血管疾病排名第二）。在25～44岁的美国人群中，1995年，艾滋病是致死的头号原因（现在已降至第六位）。死者中每年仅有2/3的人死于自然原因。那么那些情况不严重但影响深远的疾病又如何？比如普通感冒。据统计，美国民众每年总共会得10亿次感冒，平均每人3次。普通感冒是各种鼻病毒引起的，其中大约有99种已经排序，种类之多是普通感冒长久以来如此难治的根源所在。

在医疗保健方面的应用，除了分析并指出非自然死亡的原因之外，大数据同样也可以增加医疗保健的机会，提升生活质量，减少因身体素质差造成的时间和生产力损失。

以美国为例，通常一年在医疗保健上要花费27万亿美元，即人均8 650美元。随着人均寿命的增长，婴儿出生死亡率降低，更多的人患上了慢性病，并长期受其困扰。如今，因为注射疫苗的小孩增多，所以5岁以下小孩的死亡数降低。而除了非洲地区，肥胖症已成为比营养不良更严重的问题。在比尔与美琳达·盖茨基金会以及其他人资助的研究中，科学家发现，虽然世界人口寿命变长，但人们的身体素质却下降了。所有这些都表明我们急需提供更高效的医疗保健，尽可能地帮助人们跟踪并改善身体健康。

5.2.1 量化自我，关注个人健康

谷歌联合创始人谢尔盖·布林的妻子安妮·沃西基（同时也是公司的首席执行官）2006年创办了DNA[①]（图5-5）测试和数据分析公司23andMe（图5-6）。公司并非仅限于个人健康信息的收集和分析，而是将眼光放得更远，将大数据应用到了个人遗传学上。2016年6月22日，《麻省理工科技评论》评选出了50家“最智能”科技公司，23andMe排名第7。

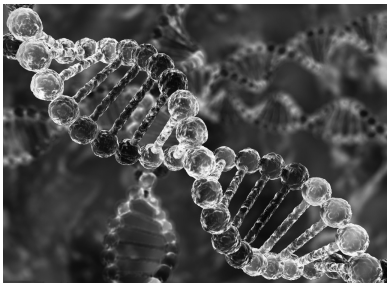

图5-5 基因DNA图片

图5-6 23andMe的DNA测试

通过分析人们的基因组数据，公司确认了个体的遗传性疾病，如帕金森氏病和肥胖症

① DNA：脱氧核糖核酸（Deoxyribonucleic acid），又称去氧核糖核酸，是一种分子，可组成遗传指令，以引导生物发育与生命机能运作。

等遗传倾向。通过收集和分析大量的人体遗传信息数据,该公司不仅希望可以识别个人遗传风险因素,以帮助人们增强体质并延年益寿,而且希望能识别更普遍的趋势。通过分析,公司已确定了约 180 个新的特征,例如所谓的“见光喷嚏反射”,即人们从阴暗处移动到阳光明媚的地方时会有打喷嚏的倾向;还有一个特征则与人们对药草、香菜的喜恶有关。

事实上,利用基因组数据来为医疗保健提供更好的洞悉是自 1990 年以来所做努力的合情合理的下一步。人类基因计划组绘制出总数约有 23 000 组的基因组,而这所有的基因组也最终构成了人类的 DNA。这一项目费时 13 年,耗资 38 亿美元。

值得一提的是,存储人类基因数据并不需要多少空间。有分析显示,人类基因存储空间仅占 20MB,和在 iPod 中存几首歌所占的空间差不多。其实随意挑选两个人,他们的 DNA 约 99.5%都完全一样。因此,通过参考人类基因组的序列,也许可以只存储那些将此序列转化为个人特有序列所必需的基因信息。

DNA 最初的序列在捕捉的高分辨率图像中显示为一列 DNA 片段。虽然个人的 DNA 信息以及最初的序列形式会占据很大空间,但是,一旦序列转化为 DNA 的 As、Cs、Gs 和 Ts,任何人的基因序列就都可以被高效地存储下来。

数据规模大并不一定能称其为大数据。真正体现大数据能量的是不仅要具备收集数据的能力,还要具备低成本分析数据的能力。虽然,人类最初的基因组序列分析耗资约 38 亿美元,不过,如今只需花大概 99 美元就能在 23andMe 网站上获取自己的 DNA 分析。业内专家认为,基因测序成本在短短 10 年内跌了几个数量级。

当然,仅有 DNA 测序不足以提升人们的健康,也需要在日常生活中做出改变。

5.2.2 可穿戴的个人健康设备

2021 年 11 月 17 日,华为正式发布了首款搭载鸿蒙操作系统的华为 Watch GT 系列新作 GT3(图 5-7)。在这个人们越来越重视身体健康的时代,相对于手机,智能手表作为可穿戴产品有着与生俱来的优势。由于可以和身体直接接触,可以更直接地帮助人们了解自己的健康状态。

图 5-7 华为 Watch GT3

华为 Watch GT3 是一款在尺寸上和交互体验上最接近传统形态的智能手表。在材质方面,前壳部分采用不锈钢材质,除了可以提升手表的质感,还可以有效避免日常生活

中的刮擦对于机身的磨损。背壳部分采用高分子的纤维复合材料，经过陶瓷效果烤漆后，有效提升后壳的防冲撞能力。背壳部分弧形的心率镜片除了可以进行相应的健康数据测试，还能有效地避免长时间佩戴手腕部分汗液的累积。弧形设计更加符合人体生物学结构，均匀分摊手腕的压力，提升佩戴舒适度。

如今，人们越来越关注自己的身体健康状况，而智能手表凭借出色的穿戴性以及专业的健康管理功能深受人们青睐。华为 Watch GT3 搭载了自主研发的第五代华为 TruSeen™心率技术，加上 8 合 1 透镜 LED 发光芯片和多路收光设计，在提升精准度的同时也降低了功耗。当佩戴在手腕上时，能够 24 小时连续进行心率检测。同时通过相应的算法优化，提升了运动中心率的准确度。

常规的健康功能还包括呼吸健康、睡眠情况、情绪压力以及女性的生理周期管理等。基于高性能的心率传感器，华为 Watch GT3 与中国 301 医院专家团队开展了联合健康研究计划。提供房颤及早搏筛查、个性化指导、房颤风险预测等整合管理服务。华为官方信息显示，截至 2021 年 9 月 30 日，共有 320＋万华为穿戴用户加入心脏健康研究，经过心脏健康研究 APP 筛查，疑似房颤 11 415 人，医院回访 4916 人，确诊 4613 人，准确率高达 93.8％。

体温是作为显示人体生命体征最为重要的一个表征。尤其是疫情期间，体温检测更是排查病例的一项重要环节。华为 Watch GT3 的高精度温度传感器能够帮助用户更好、更便捷地了解体温。华为 Watch GT3 新增了高原关爱模式，根据检测到的海拔、心率以及血氧的数据，对用户进行相应的高原反应风险评估。启动高原关爱模式（图 5-8）后，当用户的血氧水平与常规血氧有一定差距时，手表会进行相应提醒，并给出科学的呼吸调整。这对于长期处于高海拔地区或是前往高海拔地区出行的用户，还是相当实用的。

图 5-8 华为手表的高原关爱模式

美国的移动电子医疗公司 Fitbit、耐克公司、可穿戴技术商身体媒体公司（Body Media）都有手环、臂带等可穿戴的运动监测产品。

据出自美国心脏协会的文章《非活动状态的代价》称，65％的成年人不是肥胖就是超重。自 1950 年以来，久坐不动的工作岗位增加了 83％，而仅有 25％的劳动者从事的是身体活动多的工作。美国人平均每周工作 47 个小时，相比 20 年前，每年的工作时间增加了 164 个小时。而肥胖的代价就是，据估计，美国公司每年与健康相关的生产力损失高达 2 258 亿美元。因此，类似华为手表、Fitbit 手环这样的设备对不断推高医疗保健和个人健康的成本确实有影响。通过这些应用程序收集到的数据，可以了解正在发生什么以及身体状况走势怎样。比如，如果心律不齐，就表示健康状况出现了某种问题。通过分析数百万人的健康数据，科学家们可以开发更好的算法来预测人们未来的健康状况。

回溯过去，检测身体健康发展情况需要用到特殊的设备，或是不辞辛苦、花费高额就诊费去医生办公室问诊。可穿戴设备新型应用程序最引人瞩目的一面是：它们使得健康信息的检测变得更简单易行。低成本的个人健康检测程序以及相关技术甚至“唤醒”了全

民对个人健康的关注。当配备合适的软件时,低价的设备或唾手可得的智能手机可以帮助人们收集很多健康数据。将这种数据收集能力、低成本的分析、可视化云服务与大数据以及个人健康领域相结合,将在提升健康状况和减低医疗成本方面发挥巨大的潜力。

5.2.3 大数据时代的医疗信息

就算有了可穿戴设备与应用程序,人们依然需要去看医生。大量的医疗信息收集工作依然靠纸笔进行。纸笔记录的优势在于方便、快捷、成本低廉。但是,因为纸笔做的记录会分散在多处,就会导致医疗工作者难以找到患者的关键医疗信息。

2009 年颁布的美国《卫生信息技术促进经济和临床健康法案》旨在促进医疗信息技术的应用,尤其是电子健康档案的推广。法案也在 2015 年给予医疗工作者经济上的激励,鼓励他们采用电子健康档案,同时会对不采用者施以处罚。电子病历(图 5-9)是纸质记录的电子档,如今许多医生都在使用。相比之下,电子健康档案意图打造病人健康概况的普通档案,使得它能被医疗工作者轻易接触到。医生还可以使用一些新的 APP 应用程序,在平板电脑、智能手机、搭载安卓系统的设备或网页浏览器上收集病人的信息。除了可以收集过去用纸笔记录的信息外,医生们还将通过这些程序实现从语言转换到文本的听写、收集图像和视频等其他功能。

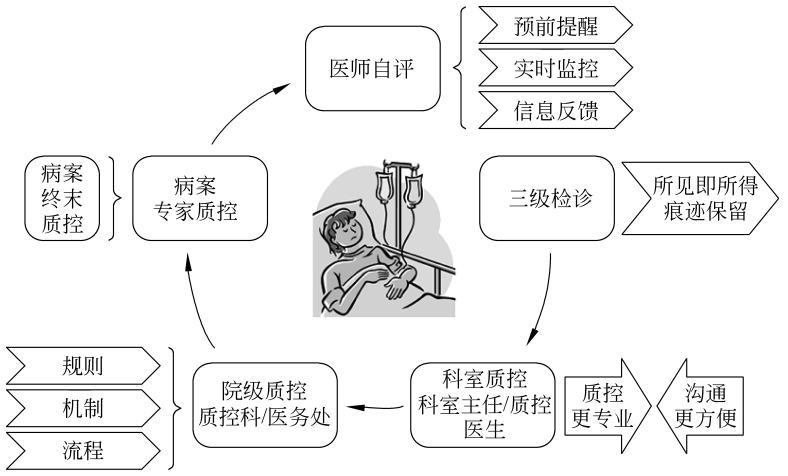

图 5-9 电子病历

电子健康档案、DNA 测试和新的成像技术在不断产生大量数据。收集和存储这些数据对于医疗工作者而言是一项挑战,也是一个机遇。不同于以往采用的封闭式的医院 IT 系统,更新、更开放的系统与数字化的病人信息相结合可以带来医疗突破。

如此种种分析也会给人们带来别样的见解。比如,智能系统可以提醒医生使用与自己通常推荐的治疗方式相关的其他治疗方式和程序。这种系统也可以告知那些忙碌无暇的医生某一领域的最新研究成果。这些系统收集、存储的数据量大得惊人。越来越多的病患数据会采用数字化形式存储。这不仅包括人们填写在健康问卷上或医生记录在表格里的数据,还包括智能手机和平板电脑等设备以及新的医疗成像系统(比如 X 光机和超

音设备)生成的数字图像。

就大数据而言,这意味着未来将会出现更好、更有效的患者看护,更为普及的自我监控以及防护性养生保健,也意味着要处理更多的数据。其中的挑战在于,要确保所收集的数据能够为医疗工作者以及个人提供重要的见解。

5.2.4 CellMiner,对抗癌症的新工具

所谓 PSA,是指前列腺特异抗原。PSA 偏高与前列腺癌症紧密相关。即使检查本身并没有显示有癌细胞,但 PSA 偏高的人通常会被诊断出患有前列腺癌。是否所有 PSA 高的人都患有癌症,这难以确诊。对此,一方面,患者可以选择不采取任何行动,但是必须得承受病症慢慢加重的心理压力,也许终有一日会遍至全身,而他已无力解决;另一方面,患者可以采取行动,比如进行一系列治疗,从激素治疗到手术切除,再到完全切除前列腺,但结果也可能更糟。选择对于患者而言,既简单又复杂。

这其中包含两个数据使用方面的重要经验教训。

(1) 数据可以帮助我们看得更深入。数据可以传送更多的相关经验,使得计算机能够预知想看的电影、想买的书籍。但是,涉及医药治疗时,通常来说,就如何处理见解这一问题,制订决策可不容易。

(2) 数据提供的见解会不断变化发展。这些见解都是基于当时的最佳数据。正如试图通过模式识别出诈骗的诈骗检测系统在基于更多数据时能配备更好的算法并实现系统优化一样,当掌握了更多的数据后,对于不同的医疗情况会有不同的推荐方案。

对男性来说,致死的癌症主要是肺癌、前列腺癌、肝癌以及大肠癌,而对于女性来说,致死的癌症主要是肺癌、乳腺癌和大肠癌。抽烟是引起肺癌的首要原因。1946 年,抽烟人数占美国人口的 45%,1993 年降至 25%,到了 2010 年降至 19.3%。但是,肺癌患者的五年生存率仅为 15%,且这一数字已经维持 40 年未变。尽管如今已经是全民抗癌,但目前仍没有癌症防治的通用方法。很大原因在于癌症并不止一种——目前已发现 200 多种不同种类的癌症。

美国国家癌症研究所隶属于美国国立卫生研究院,每年用于癌症研究的预算约为 50 亿美元。癌症研究所取得的最重大进展就是开发了一些测试,可以检测出某些癌症,比如 2004 年开发的预测结肠癌的简单血液测试。其他进展包括将癌症和某些特定病因联系在一起。比如 1954 年一项研究首次表明吸烟和肺癌有很大关联,1955 年的一项研究则表明男性荷尔蒙睾丸素会促生前列腺癌,而女性雌激素会促生乳腺癌。当然,更大的进展还是在癌症的治疗方法上。比如,发现了树突状细胞,这是提取癌症疫苗的基础;还发现了肿瘤通过生成一个血管网,为自己带来生长所需的氧气的过程。

美国国家癌症研究所癌症研究中心研制的“细胞矿工”(CellMiner)是一个基于网络形式、涵盖了上千种药物的基因组靶点信息的工具,为研究人员提供了大量的基因公式和化学复合物数据。这样的技术让癌症研究变得高效。该工具可帮助研究人员用于抗癌药物与其靶点的筛选,极大地提高了工作效率。通过药物和基因靶点的海量数据相比较,研究者可以更容易地辨别出针对不同癌细胞具有不同效果的药物。过去,处理这些数据集意味着要处理运作不便的数据库,因而,分析和汇聚数据也就异常艰难。从历史角度来

看,想用数据来解答疑问和可以接触到这些数据的人不重叠且有很大代沟。而如“细胞矿工”一样的科技正是缩小这一代沟的工具。研究者们用“细胞矿工”的前身,即一个名为“对比”(COMPARE)的程序来确认一种具备抗癌性的药物,事实证明,它确实有助于治疗一些淋巴瘤。而现在,研究者们使用“细胞矿工”弄清生物标记,以了解治疗方法有望对哪些患者起作用。

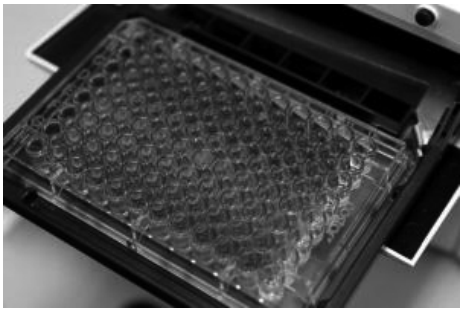

图 5-10　装载 NCI-60 细胞系的细胞板

CellMiner 软件以 60 种癌细胞为基础,其 NCI-60 细胞系是目前使用最广泛的用于抗癌药物测试的癌细胞样本群(图 5-10)。用户可以通过它查询到 NCI-60 细胞系中已确认的 22 379 个基因,以及 20 503 个已分析的化合物的数据(包括 102 种已获美国食品和药物监督局批准的药物)。

研究者认为,影响力最大的因素之一是可以更容易地接触到数据。这对于癌症研究者,或是对那些想充分利用大数据的人而言是至关重要的一课——除非收集到的大量数据可以轻易为人所用,否则他们能发挥的作用就很有限。大数据民主化,即开放数据,至关重要。

5.3　医疗信息数字化

医疗领域的循证试验已经有一百多年的历史了。早在 19 世纪 40 年代,奥地利内科医生伊格纳茨·塞麦尔维斯就在维也纳完成了一项关于产科临床的详细的统计研究。塞麦尔维斯在维也纳大学总医院首次注意到,如果住院医生从验尸房出来后马上为产妇接生,产妇死亡的概率更大。当他的同事兼好朋友杰克伯·克莱斯卡死于剖腹产时的热毒症时,塞麦尔维斯得出一个结论:孕妇分娩时的发烧具有传染性。他发现,如果诊所里的医生和护士在给每位病人看病前用含氯石灰水洗手消毒,那么死亡率就会从 12%下降到 2%。

这一最终产生病理细菌理论的惊人发现遇到了强烈的阻力,塞麦尔维斯也受到其他医生的嘲笑。他主张的一些观点缺乏科学依据,因为他没有充分解释为什么洗手会降低死亡率,医生们不相信病人的死亡是由他们所引起的,他们还抱怨每天洗好几次手会浪费他们宝贵的时间。塞麦尔维斯最终被解雇,后来他精神严重失常,并在精神病院去世,享年 47 岁。

塞麦尔维斯的死是一个悲剧,成千上万产妇不必要的死亡更是一种悲剧,不过它们都已成为历史,现在的医生当然知道卫生的重要性。然而,时至今日,医生们不愿洗手仍是一个致命的隐患。不过最重要的是,医生是否应该因为统计研究而改变自己的行为方式,至今仍颇受质疑。

唐·博威克是一名儿科医生,也是保健改良协会的会长,他鼓励进行一些大胆的对比试验。十几年以来,博威克一直致力于减少医疗事故,他也与塞麦尔维斯一样努力根据循