大数据时代的医疗革命

史今驰　著

天津出版传媒集团
天津科学技术出版社

Preface 前言

大数据的应用如今逐渐渗透在各个领域。大数据，已然成为当今时代引领各行各业变革的关键引擎。然而，在医疗领域，大数据的应用只有通过与临床，医学工程与计算机技术的跨界融合，才能真正达到落地实践服务临床的目的。如何以数据创新探索未来的医学科学，如何在庞大的数据资源中快速获取信息、提升人类医疗集体经验，是亟待探讨的现实问题。

这本书汇集了中国医学科研前沿研究成果，首先从医疗信息化历史沿革人手，结合大数据时代下的国内外医疗现状，给出医疗大数据的基本概念和变革内容，包括定义、医疗大数据技术等问题；然后针对不同场景，介绍了临床、药学、中医和公共卫生等多种大数据应用实例，解析了区域医疗和健康物联两大主题中的大数据服务问题，据此对医疗大数据的未来应用进行全景式展望。

医疗数据的共享和应用的广泛，也将面临着更多安全挑战。医院在网络安全层面遭遇了勒索病毒，勒索病毒初发时，国内医疗机构中招的很鲜见，但2018年成为局部爆发年。某些医院中了勒索病毒后引起HIS故障，有些会在微信群曝光和讨论，当然更多医院是低调处理。在数据应用方面，由于数据挖掘人员接触数据的机会增多、医院内外网业务打通的通道增多，应用风险也随之增大。

网络安全等2.0的出台以及防病毒软件的道高一尺，并不意味着2019年更加安全。唯有做到核心业务系统的隔离，采取抗感染的操作系统，尝试数据利用的去隐私和加密方式，加强数据利用审核流程，加强全员信息安全责任和管理，才能做到遇事不乱，减少乃至避免数据安全风险。

医疗大数据积累了20年，就是为医疗和科研服务的，不能再以安全为

名继续沉睡下去。借用原国家卫生计生委副主任金小桃的话说:“健康医疗大数据需要人人参与贡献,至于保护隐私和安全,随着健康医疗大数据的技术发展,这些都不是问题。”

本书主要作为医学领域从事医疗大数据研究工作者的参考书,也可以为医疗领域的相关管理人士提供参考和启发。本书在编写过程中参阅了大量是文献和资料,如有雷同,敬请批评指正。希望能对愿意参与到医疗领域大数据变革中来的读者有所启迪。

Contents 目 录

第一章　数据、信息、知识的意义

第一节　直面“大数据”

“量变会转化为质变”。

曾屹立于国际象棋世界王座不倒的象棋大师卡斯帕罗夫(GarryKasparov)在与 IBM 推出的超级计算机深蓝(Deep Blue)展开人机大战后,说了这样一番意味深长的话。

“至少在国际象棋的世界里,人类已经无法战胜拥有压倒性数据和计算能力的计算机!”

量变的确会转化为质变。

今天,随着大数据时代的到来,摆在我们眼前的课题也已经蜕变为量和质的问题。怎样正确驾驭这些以令人恐惧的势头持续增加的数据,又怎样将它们转化为高质量的信息?

毋庸置疑,各种数据的爆发式增加会为商业世界创造新的机遇,也会为医学领域的学术研究带来巨大的福音。但是,数据同时也是一柄双刃剑,读取方式错误,可能造成致命的灾难性后果。

研究人员在面对数据时,应该从以下两方面的视角出发。

这就是“假设验证”和“假设生成(一说探索)”。其中,假设验证可以说是研究人员的基本态度,为验证自己提出的假设正确与否需要收集大量数据。当数据与假设不一致,还要对假设进行修改。通过反复的“假设与验证”程序,假设逐渐凝练,继而找到通向真理的道路。

那么,假设来自哪里?

首先,是研究人员的“限定范围内的观察区域”和“大脑”。

其次,“先行研究(现存文献)”是迄今为止主要假设的生成场所。其样貌正逐步改变。从庞大的数据海洋中。或许在与人类前所未有的感知形式相左的过程中,某个新假设突然出现。日新月异的计算机分析技术,例如数据挖掘也许会承担假设萌芽的一个辅助角色。结果,某个研究人员提出的假设也许呈现出颠覆传统的奇葩内容。数据挖掘(Data mining),又译为资料探勘、数据采矿。它是数据库知识发现中的一个步骤。

以抢在竞争对手前面出招为制胜策略的商业世界为例,商业人惯常的思维模式是根据数据挖掘的结果生成某种假设,在这个无沦成败与否的假设阶段,首先果断出手,小试牛刀。

例如:市场营销界有一个著名的定论,即“来超市买啤酒的男性顾客很大概率也会购买婴儿尿布”。这是某连锁超市对庞大的历史销售数据进行研究后得出的结论。那么,据此类推,如果这家超市大胆尝试在饮料区的收银台附近摆放婴儿尿布会如何其结果,即使尿布销量平平,对这家超市的销售业绩也无妨大碍,但是,医疗的世界绝非如此。

假设大数据分析获得如下结果,证明:“人在食用某种特定食品后,会出现血压降低的情况。”当然,仅仅根据这个结果,没有哪个研究人员会得出“这种食品具有降低功效”的结论。但是,即便如此,从事人体血压波动研究的研究人员却不能轻易放过这项数据,因为从这种食品中所含的某种有效降压成分与血压之间的关系展开分析,他们可能得出某个新的假设。

对大数据进行分析,就能获得传统无法与之比拟的真知灼见,继而获得大量假设、这其中既包括新派生的某种假设,也可能是一些“无聊或不值一提”的假设、可以说,随着大数据时代的到来,医

疗从业人员最应拿出的态度，第一是“假设验证”，第二是“假设生成”的平衡。

如前所述，假设验证是接近科学的根本。但是，凡事过犹不及，如果太钻牛角尖，就可能陷入诸如“执着于寻找对自己提出的假设有利的数据，仅仅将着眼点放在这类数据上”；或者“即使客观数据无法形成佐证，也固执地生搬硬套在自己的假设上……”这样危险的境地。

让人遗憾的是，这恐怕就是近年来在社会上掀起狂澜的“科学论文抄袭问题”之类事件的背景之一。另一方面，与假设验证相比，假设生成在某种程度上需要自由思考，即所谓的“独立假设”，这对科学家来说具有极大的吸引力，但同时必须远离个人主观或自以为是的根源陷阱，以一颗虚怀若谷的坦诚之心，倾听以事实为根据的客观数据。

接下来，对这种情况下浮现出来的某个征兆进行切实的提炼，将这个雾里看花的事物作为假设的萌芽进行耐心培育，在此基础上寻找相关数据。当假设以某种形式逐渐形成雏形，此时需要做的是从验证的视角与数据进行关联——这种假设生成看似具有无穷的吸引力，实则可能以某种完全背离传统知识积累的、无法解释的关系呈现出来。大数据时代的假设生成，其优点和注意点带有这种前所未有的鲜明特征。

本节将从医疗从业者面对大数据时的正确立场出发，例如：以诱发高血压的病因为例，假设人体缺乏某种维生素会导致血压升高。那么，为了对这个假设进行验证，必须以人为对象开展流行病学研究，进行相关数据的搜集和分析工作，然后对结果进行分析（信息提取）。

结果，如果掌握以下两类数据：(1)少量摄入这种维生素的人群存在血压升高的倾向；(2)服用这种维生素的人群的血压呈下降趋势，那么，就可以认为该假设暂时成立。

这种假设验证的思维模式不仅限于维生素与血压之间的关系，研究人员应保持“哪些物质会引起血压升高”的原始问题意识，以及与数据对立是假设生成的立场，只要对诊疗明细数据和电子病历进行充分分析，我们就会发现很多以前忽视的患者的生活方式、既往病史以及以往经历等问题。而关注这类问题有时可能与血压的新研究发展之间存在某种关联性——但是，假设生成看似具有无穷的吸引力，如果非要从否定的立场出发看问题，那么，也有可能与传统知识积累相左，陷入某种难以解释的关系之中。

大数据的世界同样可能夹杂着“偶然性”，这一点毋庸置疑，今天，即使通过大数据分析，得出两种事物之间的关联存在统计学上的意义，也不能排除在其他时间段发现同样关联关系的可能（大数据随着时间流逝，其数据的蓄积量递增，继而形式发生能动性的改变），有时，两种事物之间最初呈现的关系可能仅仅是一个“偶然”。

第二节　信息来自数据分析

数据、信息、知识究竟是什么？我们应该怎样理解它们之间的区别和关系呢？

先说数据。一般情况下，提到数据，人们第一个想到的数字，其次是文字数据。此外，声音数据、图像数据是近年来普及的概念。那么，“数据”，这个我们平时经常无意识地使用的词汇到底是什么意思？

辞典对“数据”一词的表述如下：“数据，指用于立论、计算基础的现有或者被认可的事实和数值”（引自《广辞苑》第五版）。此外，数据一词，还包括与以下分析相对的下述定义。

“以文字、符号、数值等综合形式对某种事物进行再现的结果；对人类有价值的事物；人类将分析结果称为信息的事物。单纯提及数据时，在IT业特指计算机记录、处理的内容。此外，指计算机存储中程序以外的内容以及程序处理对象等。”

例如，下面有一组医疗数值数据。

110、120、130、140……

如果为上述数据补充一些说明性的文字数据，比如“患者A－150、患者B－130……”，我们就能从中看到一些有用的信息，继而推测这组数据可能是一些代表血压的数字，也可能是一组代表小学生身高的数据。这时，如果在此基础上进一步补充诸如“患者A：收缩压150mmhg/50岁；患者B，收缩压130mmhg/35岁……”等文字数据，这组数据代表的含义将变得更加清晰。如果再将这组数值数据与文字数据并列起来看，就可以得出如“人随着年龄增加，收缩压会随之增加……”即年龄与血压之间关联关系的结论。

也就是说，当数据单独存在时，我们很难从数据中获取“意义”。相反，为了让数据变得有“意义”，必须与其他数据组合。通过在初始数字数据上追加“血压”数据，数据才变得更具“意义”。

这种从数据中产生的“意义”就是信息。辞典对“信息”一词的表述如下：“以判断或引发行动为目的，需要借助各种必要媒介的知识。”（引自《广辞苑》第五版）此外，信息的另一个定义是“通过文字、声音等途径的再现促使人在感知过程中唤起某种意义，并对具体的思考以及行为方式产生影响的事物。这与对人而言无意义的杂音以及随机模式在内的数据存在本质上的区别”。

总之，可以认为“对人有无意义”是信息与数据的区别所在，或者可以说“数据分析的结果”就是信息。那么，接下来需要我们思考的问题是怎样对数据进行分析，即数据分析的方法。

假设下面有一组数据：某人的血压值为“150/96mmHg”。

那么，从这组数据能够获得什么信息呢？根据数据读取意义，首先，数据本身的正确性是大前提。使用有故障的血压计测量出来的血压值没有任何意义。而且，隔着厚重的衣服测量出来的血压值也不可能准确。换句话说，使用精准度高的血压计，采用正确的测压方法测得的血压值才是进行数据分析的前提条件。

假设“150/96mmHg”是测量得出的正确的血压值。那么，我们从这个数据能够读取哪些信息呢？一般情况下，血压值高于“140/90mmHg”时，临床上会被诊断为高血压。按标准值衡量，根据“150/96mmHg”这个数据，可能得出这名患者患有高血压这一信息。

但是，一次测量结果还不能确诊患者一定患有高血压。即使真的患有高血压，治疗的第一步也不是立即让患者服药，而是从改善生活方式开始。

也就是说，对这个“150/96mmHg”的血压数据进行分析，为了让生成的信息为做出一个判断提供帮助，该判断标准的周边必须有一些必备知识作为基础支持。

如前所述，有时数据会呈现出某些偶然的关联关系。研究人员不能只求速度，不求精度地急着从数据中寻找信息，而应在“这种现象可能仅仅是一个偶然”等（统计学）知识支持的基础上，抱着“其他数据是否存在同样的倾向”或“是否从大量数据分析结果中提取所需数据”等谨慎的态度进行数据分析。综上所述，首先，对数据进行正确测量是数据分析的前提条件。其次，将正确数据作为知识运用并进行正确的分析。只有这样，我们才能获得恰当的信息。

第三节　数据分析的方向性和风险

一般来说，进行数据分析有两个基本视角。第一，如前文血压事例所示，需要观察与其他数据之间的关联性。单纯从数据本身无法获取任何有价值的信息。而怎样设置关联的数据项，将在很大程度上改变数据分析的难度。

第二，观察数据的分布情况。从平均值开始，中央值、标准偏差等数据之间的差异，或根据缺失值的类型读取的信息就是数据分析的根本。数据分布是今后处理大数据时最重要的衡量指标，这一点不变。

两种数据的分布方式之间存在关联性。也就是说，自变量增长时，因变量也随之增长的现象被称为“正相关”（两个变量变动方向相同，一个变量由大 Nd，变化时，另一个变量亦由大 NII，）。另一方面，因变量值随自变量值的增大而减小的现象被称为“负相关”。

在传统医疗世界中，可利用数据的绝对量是有限的。例如，在基础医学领域，实验室范围内收集的数据最多只有数十例。

在此基础上，以按比率增加（scale—up）的临床研究为例，充其量不过数百例规模；即使以地区对对象进行的流行病学研究，一个研究小组处理的数据量按受试者人数确定，通常也不过数千单位。

在大数据时代，上述可利用数据的绝对量将一跃膨胀至十万单位甚至百万单位。随着电子病历数据和诊疗明细数据进一步集约化，研究人员面对的可利用数据量很容易突破百万单位。

另一方面，人们在数据分析上花费的时间将大幅缩短。过去，当研究人员面对庞大的数据量进行复杂的统计分析时，往往需要花费很长时间才能得到结果。而现在，在计算机硬件和数据处理软件的帮助下，研究人员在处理大容量数据上耗费的时间已经显著缩短。

这里有一点需要注意：这就是从事数据处理的人——通常意义上被称为“数据科学家”这一职业的存在。

不具备医学及医疗专业知识的人担任数据科学家时，很容易发生以下问题。

假设以某国的女性为对象，围绕吸烟对人体健康的影响展开调查研究。调查开始时，以是否吸烟作为衡量指标，将调查对象分为两组，在接下来的 20 年期间，分别对两个调查组的生存状况展开跟踪调查；该研究的定性是流行病学染色体组型分析研究。调查数据证实：调查时吸烟组女性 20 年期间的死亡率为 13%，非吸烟组的死亡率为 19%。两组调查开始的人数均为 1 万人。

根据该数据，数据科学家得出结论：“吸烟女性比非吸烟女性的死亡率低，即长期生存率高。”可想而知，即使数据科学家坚持上述观点，也不会有哪个医学研究人员会表示支持。因为从科学常识的角度来说，“一般情况下，与非吸烟者相比，吸烟者的寿命更短”。

这组数据存在的问题是调查开始时是否追加年龄项？

如果将吸烟组和非吸烟组按 65 岁年龄段分为两组。那么，65 岁以下的吸烟者是 9000 人，非吸烟者 1000 人，死亡人数分别是 900 人和 50 人（死亡风险分别是 0.1 和 0.05）；65 岁以上的吸烟者是 1000 人，非吸烟者 9000 人，死亡人数分别是 400 人和 1800 人（死亡风险分别是 0.4 和 0.2）。

也就是说，无论年龄超过 65 岁还是未满 65 岁，吸烟者的死亡风险均为非吸烟者的 2 倍。但

是，如果不分年龄段对数据进行汇总处理，得出的结果就是吸烟人群及非吸烟人群各 1 万人，死亡人数分别为 1300 人和 1850 人，死亡风险分别为 13％和 19％。

那么，原本已成定论的“吸烟会增加死亡风险”这一关联，由于未考虑年龄因素，结果得出截然相反的结论。年龄是衡量死亡风险的重要背景因素，进行数据分析时，缺乏这方面专业知识的数据科学家很容易在根本性的问题上犯错。

如上所述，当不具备医学及医疗专业知识的数据科学家进行数据分析时，存在一个死穴——在必要的基础数据上犯错。在上述案例中，他们所犯的错误是遗漏了流行病学上重要的混淆变量(Confounding Variable)——年龄因素。当数据量增至 10 万以上时，调查项目中必须追加年龄段的吸烟/非吸烟，饮酒/不饮酒、未婚/已婚、大学毕业/非大学毕业……等多领域的指标。

当对上述数据进行分析的数据科学家得出诸如：“非吸烟、未婚、大学学历女性的死亡率最高”的信息时，我们应该怎样看待这个结果呢？

当然，必须考虑存在某种混淆变量时的情况。例如，大学毕业、未婚、不吸烟、不饮酒的女性多数埋头努力工作，与普通女性相比，她们无论在精神或体力层面都承受着更大的压力，这一点不可否认。当然，或许数据科学家的分析可能提出某个新假设。但是，由于数据科学家缺乏基本的医学及医疗常识，其分析很可能存在某些缺陷。最糟的情况莫过于这种数据科学家可能出于某种理由，故意歪曲数据。

近年来，作为新兴的热门职业之一——数据科学家正受到社会前所未有的关注。如果从事这种职业的人接触医疗及医学行业数据，从客观上来说，他们必须具备相关领域的基础知识。当然，没有哪个人能够全面掌握医疗及医学领域庞大的基础知识。但是，他们可以通过与具备这方面专业知识的人之间的密切合作和沟通达到目的。其次，他们需要学习关于科学的公正和营私舞弊行为等方面的知识，具备职业素养的高度伦理观。关于这些人才取得社会认可的资质或标准，今后有必要进一步展开探讨。

第四节　信息的作用是减少不确定性

现代社会是信息化社会。随着网络的普及，在世界范围内流通的信息量(数据量的提法可能更恰当)正呈爆炸式增加。

以日本为例，2012 年月均网络流通数据量约为 2 300PB(拍字节或千 T 字节)，大约相当 5.7 亿张 DVD。这么说可能让人一下子有点蒙，总之现代人就生活在这个似乎永远看不到尽头的庞大信息量的包围下，这一点毋庸置疑。

辞典对“信息”一词的定义参见前文所述。这里想介绍数字理论之父——克劳德·艾尔伍德·香农(Claude ElwoodShannon)提出的定义。开辟通往信息化社会之路的香农对信息是这样定义的：“所谓信息，是用来消除不确定性的东西。”

例如，假设有人准备从 A 地前往 B 地。在没有确切信息指引的情况下，这个人到达目的地的可能性将发生变化，这一点很容易理解。首先，假设这个人出发的道路有两条，一条在左，一条在右。那么，他该选择哪条路呢？在没有任何参考线索的情况下，他选择正确道路的概率仅为 1/2，即 50％。此时，假设道路的右侧竖着一块路标，路标上清晰地标明“前往 B 地方向”，那么，这个人选择正确道路的可能性会大大提高。像这样，我们将增加到达目的地概率(消除不确定性)的东西

称为“信息”。

消除不确定性的信息在医疗领域起着决定性的重要作用。早在此之前，加拿大籍内科医生威廉·奥斯勒(William Osler)对医学是这样定义的。

“Medicine is a science of uncertainty and an art ofprobability,”“医学是一门不确定性的科学和可能性的艺术。”

奥斯勒医生主张医疗具有科学和艺术的两面性。为什么？因为医疗行为本身带有不确定性。以未破裂颅内动脉瘤为例，假设某临床医生发现某患者患有早期脑动脉瘤，需要考虑是否手术。如果不及时手术，可能引发脑动脉瘤破裂出血。那么，这种情况下需要对该患者施行预防性手术。相反，另一种情况是脑动脉瘤破裂的可能性较低，不建议施行风险性手术。手术做与不做，患者得到的信息，是医生在结合个人知识储备的基础上做出的判断，除此之外没有其他途径。

结合前文的例子：某人从A地前往B地的路标(信息)发挥的重要作用，这样我们就能理解医疗领域信息的重要性。以减少医疗的不确定性为目的的信息所承担的作用多么重要！

医疗一线充当路标作用的对象是诊疗手册。所谓诊疗手册，就是将与疾病相关的各种复杂的研究成果(evidence)进行汇总整理，为医生和患者做判断提供帮助的文档。根据美国医学研究所(Institute of Medicine)的传统定义，“诊疗手册是在特定临床状况下，为做出合理的判断，以辅助临床从业人员和患者为目的，系统制作的文档资料”。上述定义需要注意一点——即在临床从业人员之外，患者以支持辅助为目的的参与。

1997年，日本厚生省(现更名厚生劳动省)公布《日本医疗技术评价状况研讨会》报告，该报告首次将循证医学(EBM，Evidence—Based Medicine)的观点作为医疗基础纳入其中。在之后1999年公布的报告中，日本厚生省在将EBM纳入诊疗手册制作的同时，还明确提出这一概念的重要性。此后，“遵循证据的诊疗手册”制作正式启动。

目前，日本国内发行的诊疗手册数量众多，但是，现实情况下仍然存在一些使用人和使用状况不明的现象。制作合理的诊疗手册，并在临床一线合理使用，只有这样，医疗品质才有望改善。

因此，对诊疗手册制作合理与否进行评估的手段之一是AGREE工具(临床实践指南质量评估审查工具，Appraisal ofGuidelines for Research and Evaluation)。AGREE工具由23项审查项目和综合评估构成，对保证临床诊疗手册的质量会起到很大的帮助。该方法的检查内容并非诊疗手册的医学内容，而是一种用于描述诊疗手册制作过程的主数据。所谓主数据，是指“信息的信息(与信息本身相关的信息)”。所谓诊疗手册，是指哪些人，面向哪些对象，从何种目的出发制作的手册？资金提供方是谁？在什么时间？在什么证据的基础上制作等信息。通过关注这些内容来获得判断诊疗手册质量的线索。

另一方面，为了正确理解医疗信息，还需要信息受众的信息收集能力(Literacy)。与健康、医疗相关的信息应用能力被称为“健康认知力”(Health literacy)。健康认知力的代表研究学者Don Nutbeam教授(英国南安普顿大学校长)将健康素养分为三个层次：基本/技能素养、交流/沟通素养和批判素养。

技能素养，指理解文字层面的处方。在国外，以移民为中心，无法完全使用移民国语言的人群大量存在，日本民众的识字率堪称世界第一，所以，这方面的问题并不大，但是，在现实情况下，晦涩难懂的医疗专业用语被曲解误读的情况屡屡发生。

沟通素养，是在读取文字的基础上理解内容和含义，并与他人进行沟通的能力。最后，批判素养是指按自己的方式消化理解的内容，并反映在自律行为上的能力。

第五节　正确数据分析时的注意事项

到这里，本书围绕正确进行数据测定，对测定结果进行分析，到获取信息，以消除不确定性为目的应用所获取的信息展开了阐述。那么，我们应该怎样看待信息与知识之间的关系呢？信息本身可以单独存在，但是，如果单独存在的信息在需要的时候不能获得，那么，就可能因为遗漏这部分关键信息导致整体判断失误并引发错误的行动。

举个例子，假设现在有观点称“喝黑醋能减轻体重”。一般来说，人们知道很多类似的信息，比如有效疗法及各种民间疗法、营养辅助食品、名医及知名医院推荐、患者分享自己的治愈经历和心得……随着网络的广泛普及，网上随处可见这类断章取义式的信息片段。但是，这些信息中，到底哪些是真哪些是假？这就需要人们擦亮眼睛分辨是非对错。研究人员，特别是从事人类健康与疾病研究的人通常采用以下一些系统性的方法——比如对照组对比、偏倚(Bias)排除、分子分母意识、考虑混淆因素、因果联系逆转的可能性等等。这就是流行病学中以人为对象接近科学研究的途径。

医疗实践活动中积累的重要临床信息包括病例报告和病例收集(大量病历报告)。但是，仅仅根据病例报告和病例收集还不足以得出一般论的信息，因为临床上报告的病例本身从患病个体与整体病例对比出发，其产生的选择偏倚很大。而且，因为单纯的病例报告和收集缺乏对照组(control)，所以，治疗的有效性和风险因素探讨之间存在很大的局限性。病例报告对学习临床基础知识而言的重要性不言而喻。但是，其目的不是为了获得一般性的结论，而是帮助临床医生从较少的(珍贵的)病例中加深个人临床能力。作为与治疗和风险因素相关的一般论，当我们接受某种信息时，必须持有对照组意识，这一点非常重要。

解读信息，在此基础上，下一个关键点是尽量排除偏倚，代表性的偏倚分为选择偏倚、测量偏倚、混杂偏倚三类。关于选择偏倚。有这样一则逸事：

1936 年，美国进行总统选举期间，《文学摘要》(％e literaryDigest)在选举开始前以 237 万人为对象进行了民意测验，预测共和党候选人将胜出。另一方面，盖洛普公司(Gallup，盖洛普公司由美国著名的社会科学家乔治·盖洛普博士于 1935 年创立，是全球知名的民意测验和商业调查、咨询公司)仅利用一个大约 2 万人的样本，预测民主党候选人将胜出。抽样调查对象数分别是 237 万人和 2 万人，二者相差约 100 倍。

结果怎样呢？盖洛普公司正确地预测了民主党候选人的胜利。这就是大数据时代在信息真伪之辨的基础上，一个很有说服性的例子。The literary Digest 公司从本社杂志的读者、电话簿、汽车注册会员名录中选取了 1000 万人作为问卷调查对象，其中 240 万人寄回答案。

另一方面，盖洛普公司结合与样本居住地、年龄、性别、人种等与选举权人整体分布情况类似的立场出发，重点筛选了一个大约 2 万人的调查样本。结果，尽管调查对象只有 2 万人，与 The literary Digest 公司发生偏倚的 237 万人的信息相比，准确地反映出“美国民众的代表性意见”。

这个案例告诉我们：偏倚少的样本的重要性。人们通常容易陷入的思维误区之一是只要对大样本进行调查取样并进行统计分析，就能得出正确的结果。但是，这种方式可能使数据产生偏倚，引导错误的信息，直至发生致命性的错误。这就是 Theliterary Digest 公司为我们揭示的道理。这里需要注意的问题是“随机采样”(random sampling)——即对采样对象整体进行编号，再以相同的

概率随机采取子样，代替对整体进行调查的方法。以采样对象整体的缩小型（miniature）为对象进行抽样调查，通过这种方式对整体状况进行预测。可以说，随机采样是一种科学的、划时代的方法。

那么，大数据时代会发生什么呢？用极端的话说：大数据的目标是整体调查。暂且把成本面的可行性放在一边不谈，预测候选人时，以选民整体为对象实施的调查是大数据时代的常规做法。暂且撇开预测候选人不说，如果我们能够获得足够数量的电子病例和诊疗信息。那么，通过实施整体医疗调查，就不一一看，也包括以下内容。

测量偏倚、要因预知偏倚、记忆偏倚、思维偏倚、家族信息偏倚、不认可偏倚、迎合偏倚、期待偏倚、谦逊偏倚、举止偏倚、面试者偏倚……以上每一种偏倚都可能在人无意识的状态下呈现。也就是说，人们不可能彻底脱离偏倚状态。即使是以整体调查为对象的大数据，也解决不了这个问题。

其次，所谓分子分母意识，指与传统相比接近流行病学过程中重视的一点——即母本意识。例如：假设某种治疗方法对某种疾病有效。那么，即使这种治疗方法对接受治疗的 1 万名患者显示效果，仅仅根据这一结果，也不能断言这种治疗方法一定有效。为什么？因为在 1 万名治疗有效的患者背后，可能还有 9 万名患者接受同样的治疗后无效。这被称为“脱落例”，这是解读信息时人们最容易陷入的陷阱。为了防止落入此类陷阱，研究人员必须对受试者整体进行调查。因此，如果分子分母问题能够实现与全数调查接近的大数据调查，消除偏倚的可能性无疑将大大提高。

在因果关系验证阶段，还需要结合混淆因子进行思考。这里列举一个容易把人绕进去的例子——“运动可以缓解（不患）感冒”。据美国某大学发起的问卷调查结果证实：与每天只运动 1 小时的人相比，每天进行 3 小时中等强度运动的男性患感冒的概率降低 35%。根据这项结果，就能断定“运动可以不患感冒”吗？

即使运动与不患感冒之间存在关联，那么，运动以外的其他因素也可能让人不患感冒。例如：我们可以认为经常运动的人有充足的时间用来锻炼身体，这时，不仅运动能够帮助人体远离感冒，优越的经济保障、良好的生活环境、均衡的营养状态、接受的教育程度高等因素都有可能帮助人体预防感冒。其中，特别是作为混淆因子的社会经济层面的因素是无法忽视的存在。

这个案例为我们揭示了“因果逆转”的现象。例如：假设这项调查基于以下两个问题展开，一是“最近三个月的运动程度怎样？”，二是“最近三个月是否感冒，感冒了几次？”那么，原因与结果之间的关系就可能发生逆转。也就是说，因为没有感冒，所以锻炼身体。这种“因果逆转”现象在同一时间段采取原因因素和结果因素的“横断研究”过程中尤其需要注意。

当然，病例报告和少量病例收集对医学进步发挥着巨大的作用。特别是在进行新病种发现和罕见病治疗方法研究时，这类研究不可或缺。今后，在临床一线重视病例报告这一传统的同时，临床流行病学信息的应用及普及，推动大数据运用等状况无疑将发挥巨大作用。

第六节　知识是信息结构化的总体

所谓知识，可以说是通过数据分析产生的结构化和体系化的信息。与某领域相关的这种知识集合——即知识体系（Bodyof Knowledge）是专家们的“教科书”。所谓教科书，指专业知识的结构化、相互关联以及系统记述的存在。

另一方面，在医学世界中，作为与这种教科书并列的重要文档的存在，如诊疗指南。关于诊疗指南的定义如下。

“对于诊疗过程中重要度高的医疗行为，进行有证据的系统评价（systematic review）及其系统评估，从利弊平衡等角度出发，以支持医患主张决定为目的的最佳推荐文档资料。”（《诊疗指南制作入门 2014》，日本医疗机能评估机构）如上所述，诊疗指南基础的重点是通过系统评价重视证据体系（Body of Evidence）评估。例如，假设某种疾病的治疗方法分为 AB 两种。那么，怎样判断其各自的有效件呢？

某临床研究提供的证据是三年前发表的一篇论文。在这篇论文中，指出治疗方法 A 有效。另外，某篇于一年前发表的论文报告提出治疗方法 B 有效。同时发现还有其他论文，研究人员基于文献数据库对这些论文进行系统调查后发现：这些论文共计十篇。其中七篇主张治疗方法 A 有效，三篇主张治疗方法 B 有效。那么，我们能简单地根据上述结果断定治疗方法 A 更好吗？答案当然不是。我们必须根据报告的研究规模和论文质量，结合各种事实证据进行系统评价。

假设主张治疗方法 A 有效的七篇论文均为少数人的研究成果，论文中应报告的研究内容缺失。如果论文记述不充分，那么，这项研究成果在多大程度上具有可信性呢？答案是无从判断。另一方面，尽管主张治疗方法 B 有效的论文只有三篇，但均为以多数人为调查对象实施随机采样化比照实验的结果，而且必要的佐证一项不落，内容经得起推敲和评价。如果从综合角度对上述结果进行判断，那么，最终结论是：尽管支持的论文数少，与治疗方法 A 相比，治疗方法 B 也是有效的。

目前日本国内能够实施这种系统评价的研究者人数可能只有两位数。随着大数据的普及。社会以及医疗从业人员必将面对和接触新次元的信息和知识。大数据的分析结果以及导出的信息未必都是正确的。通过对庞大的数据进行分析，不排除一些偶然发现的相关关系公开发布和报告的可能性。

系统评价是以临床试验（流行病学上的介入性研究）的证据为中心发展起来的科学方法，但是，面对大数据分析（观察研究）结果时，仍需要进行慎重的信息评估和整合作业。在迎来大数据时代的今天，我们强烈期待与进行数据直接分析的数据科学家一起，进行新的系统评价方法论研究和推动具备相关知识与技能的人才培养工作。

医疗从业人员面对数据、信息和知识时，需要具备专业水准（professionalism）和诚信（integrity）的态度。所谓专业水准，特别是临床人员与患者一起在朝着理想方向努力的同时，需要通过维护公共健康承担起相应的社会责任。近年来，人们正将怎样在有限的医疗资源下实施上述工作作为新要务展开积极探讨。

另一方面，研究者必须坚守诚信原则——通过公正的科学活动承担社会责任。社会责任认识是从事医疗大数据的研究人员应有的共同意识。

对大数据的潜在能力展开想象，那么，不仅临床从业者和研究人员，行政机关以及企业人也应具有同样的自戒和自问意识。

第二章　什么是医疗大数据

第一节　从茫茫数据洪流中吸取什么?

所谓大数据,是指数据量庞大的数据,多数来自互联网上流通的庞大的非结构化数据。从目前来看,尚无明确的关于大数据的统一定义。

2007年,日本滋贺县长滨市携手京都大学研究院医学研究科共同发起一项名为"长滨零级预防队列研究"的项目。

预防医学原本分为一次预防、二次预防、三次预防这三个阶段。一次预防是对疾病发生本身的预防,具体指生活习惯及生活环境改善等整体健康促进措施,作为特定病种开展预防接种等工作。二次预防是通过早期发现早期治疗,对疾病发展的预防。该阶段的工作是进行健康诊断,接受精密体检。日本精密体检是指以不管有无自觉症状,定期地通过CT、MRI、超声波等精密仪器,对身体各部位进行的一种精密检查。精密体检是日本最标准的癌症检查手段,每年大约有200多万人接受检查。其费用除却医保对象,原则上自费。三次预防是通过治疗防止疾病重症化,通过保健指导和康复训练等促进机体功能恢复,帮助患者回归社会,防止疾病复发的预防工作。

这些在传统预防医学基础上新增的先进措施就是"零级预防"。用一句话对零级预防进行概括——"以了解个人遗传基因的特性为目的进行代谢组学分析,根据结果采取个别性高的预防工作"。

这种措施也被称为"先制医疗"(Preemptive medicine)。此外,将基因组学(Genomics)、代谢组学(Metabolomics)、蛋白组学(Proteinomics)三个词的词尾"omics"合在一起,称"组学"(Omics)。具体地讲,首先让患者了解进行基因分析的自身体质,对未来容易罹患的疾病进行预测。根据该预测结果,从早期阶段开始改善个人生活方式、饮食,积极进行体育锻炼。通过上述方式防病于未然。从未病阶段开始采取预防措施,这被定义为"零级预防"。

研究成果将以重视项目参加者个人信息保护为前提,在一定的前提条件下面向国内外公开,同时用于推动前沿医疗研发工作。2014年,由日本独立行政法人科学技术振兴机构(JST)运营的日本国家科学数据中心(NBDC,National BioscienceDatabase Center)开始对外提供包括基因分析信息在内的数据。

"长滨零级预防队列研究项目"以长滨市在住的30～74岁之间的1万多名居民为对象,对受试者进行健康诊断、血液检查、大动脉波速、呼吸功能等约700项生理学检查。

结果,每名受试者的数据项目多达800项。如果将这些项目放在计算软件中制作成表格,横列800项,纵行记录人数(1万人),可以制成一张800×10 000项的EXCEL数据表。该项目的性质是队列研究,而不是一年期的短期调查活动。今后,将在长达十年的时间里,持续对受试者的健康状态、患病情况、寿命等进行跟踪调查。随着软件升级换代,表格计算软件、EXCEL的功能将日臻完善,所处理的数量也将呈飞跃增加。EXCEL2010开始的版本最大行数是1048576行,最大列数是16384列——即可容纳的项目多达100万人,2万项,这里提到的"长滨零级预防队列研究项目"数据不包括基因组数据。如果在此基础上增加受试者的基因组数据,每名受试者涉及的数据量将增

加 30 亿碱基对数据。在以多组人群为对象进行的流行病学研究中，将通过与基因科学的融合，在医学与医疗世界中成为推动大数据活动的重要推手。那么，我们能从大数据中获取什么呢？

首先，是对脑梗死、心肌梗死、代谢症候群为中心的生活方式病发病机理展开分析，同时，更深层次的社会问题——与长期看护及护理问题相关的老年痴呆症以及“运动障碍综合征”（随年龄增长和生活方式等伴发的骨骼、关节、肌肉等运动机能器官的非正常状态）也是一大目标。在发挥上述预防医学作用的基础上，怎样利用有限的医疗资源？怎样实现更高质量和安全性的医疗活动？大数据还将在这些领域发挥重要作用。

可以说，医疗大数据将为提高患者的生存质量（Quality ofLife）做出巨大贡献。

第二节　呈爆炸式增长的全球数据存储量

也就是说，“大数据”的意义不仅仅是与医疗行业相关的数据，甚至可以说，医疗数据在大数据中仅占很少的一部分。那么，现在全球范围内的“大数据”到底是一种什么样的状态呢？

大数据呈现爆炸式增长，从 2005 年开始进入 Web2.0 时代，网络世界的数据量进一步激增。

那么，“Web2.0”与传统模式——“Web1.0”之间的区别是什么呢？

二者之间最大的区别之一是制作互联网流通内容的难易程度。Web1.0 时代的网页语言代码 HTML 是由少数编辑人员定制的。市面上有在售的专业网页制作软件，但使用这些软件需要具备相应的专业技能，不是什么人都能简单操作的。

但是，随着 Web2.0 时代的到来，情况发生了巨大改变。Web2.0 包含了我们经常使用到的服务，比如博客、电子公告牌、图片及动画分享服务……可以说，我们每个人都是网页内容的供稿者，用户只需敲击键盘输入就能轻松推送信息，或者将数字数据、视频等上传到网上，在世界范围内公开发布自己拍摄的影像作品。互联网信息发布的难度大大降低。结果，数量庞大的用户群开始成为网络内容的作者，互联网上公开的信息量开始急剧增加。用户和消费者自身创造信启、的“ConsumerGenerated Contents”（定制生成内容）时代拉开序幕。

但是，这些由“如雨后春笋般冒出的内容制作者们”制作的内容与 web1.0 时代由专业人员制作的内容迥然不同，其质量千差万别，鱼龙混杂。互联网上的数据量暴增本身是一件好事，但是，如果按这种状态持续发展下去，哪些信息是有用的？哪些信息可信？用户根本无法甄别，更不知道在哪里能够找到所需的信息。部分信息学家将这种状况定义为“DRIP”——DataRich Information Poor（数据很丰满，信息很骨感）。

于是，一些企业在这种混沌未分的状况下迅速发展壮大，例如为用户提供免费信息检索服务的谷歌公司。1998 年由同公司创办的 Google's mission 对其搜索引擎是这样描述的。

“谷歌的使命是整合全球范围的信息，使人人皆可访问并从中受益。”

大数据时代的操盘手不止一家企业，但是，谷歌无疑是其中最重要的一家。可以说，谷歌推出的搜索引擎改变了互联网的使用方式。该公司通过引进被称为“网页排名（PageRank）”的十进制计算技术，帮助每个用户通过简单检索轻松找到所需的信息。这些信息不仅限于大多数用户寻找的一般性信息，还包括少数人关心的，希望获取的罕有信息，甚至小众信息。用户个人推送的信息可能被每个人都看到，这对信息制作者来说将激发其产生信息的热情，并进一步将制作的内容上传