大数据思维与传统统计思维差异的思考

陈 超,沈思鹏,赵 杨,陈 峰,魏永越
(南京医科大学公共卫生学院,江苏 南京 211166)

摘 要:大数据时代已经来临,随之而来的巨量信息正变革人们的思维模式以及理解世界的方式。认识到大数据思维和传统统计思维的差异,有助于思维的变革和方法论的演进,以便于充分利用数据的潜在价值,对于创新统计思维模式有重要意义。文章就大数据思维和传统统计思维的差异进行分析和探讨。

关键词:大数据思维;生物医学研究;数据挖掘;传统统计思维
中图分类号:G633.6 **文献标志码**:A **文章编号**:1671-0479(2016)06-477-003
doi:10.7655/NYDXBSS20160615

随着信息时代的飞速发展,大数据成为了人们热议的话题。大数据时代意味着这个世界正实时产生大量多源的、纷繁的结构化和非结构化信息,这些信息正影响着人类的思维及生活方式。在生物医药领域也不例外,大数据贯穿于基础研究、药物开发临床诊疗以及健康管理的所有环节[1]。由于大数据具备4V(volume,variety,velocity 和 veracity)的特点[2],这给统计分析方法提出了巨大的挑战,大数据思维亦应运而生。大数据思维与传统统计思维类似,都是对真实世界数据进行正确描述和科学分析,从而揭示事物的本质,并把握其发展变化的规律[3]。然而,由于大数据自身的特点,其思维与传统统计思维又有着本质的差异。

一、研究目的不同

在传统统计工作中,尤其是在生物医学研究中,确证性研究是长期以来的主要目的之一。基于事物间的相关性、先验信息,以及统计推断方法,进行因果关系的初步研究。但是,由于样本数据的不完整性,仍需要大量的工作进行后续的因果关系验证。在大数据背景下,并不需要了解事物发展的因果关系。大数据主要应用于探索性研究,其主要核心是建立在相关关系之上,排除人为假设,挖掘数据深处的意义,获得更多的认知与洞见,进而可以科学地预测。由于立足于总体(大数据),我们可以观察到以往注意不到的联系及很难理解的复杂现象。生物医学大数据可以广泛应用于疾病的早期检测和诊断。例如,谷歌利用搜索引擎检索词条,预测某些区域的流感传播情况。检索词条与流感传播并无明确的因果关系,但通过相关关系的研究,却能很好地预测疾病传播情况[4]。

需要指出的是,大数据思维并非完全否定因果性。对生物医学研究来说,因果关系永远是探究疾病病因的密钥。因果关系也是一种特殊的相关关系,基于大数据中所反映的相关关系,我们可以继续发掘更深层次的因果关系。

二、研究对象不同

总体性和样本性差异,可以说是大数据思维和传统统计思维最本质的差异。在传统统计中,随机抽样一直被公认为是最有效的数据收集方法。统计学家也已经证明:抽样分析的准确性随着抽样随机性的增加而大幅提高,因此样本选择的随机性比样本量更重要[5]。用小数据去窥探全体样本的面貌,是小

基金项目:国家自然科学基金项目"基于生物学调控网络的肺癌多平台组学数据的整合分析方法研究"(81402764);江苏高校品牌专业建设工程资助项目(PPZY2015A067);江苏省高效优势学科建设工程资助项目
收稿日期:2016-09-19
作者简介:陈超(1994-),男,江苏常州人,本科生在读;魏永武(1983-),男,江苏盐城人,讲师,研究方向为医疗大数据的分析方法与技术,通信作者。

数据时代处理分析数据的一条捷径。但是现实中的生物医学研究,实现从总体中进行绝对随机抽样相当困难,无法展示事物全貌,调查结果可能缺乏延展性。例如,一项生化指标与帕金森病患病风险的相关性研究,纳入自2009年1月至2013年12月在江苏省人民医院确诊为帕金森病的277例患者,以及277例年龄与性别相匹配的同期健康体检者进行此项研究[6]。其抽样的代表性不得而知,因而其结果的延展性有待验证。

在大数据时代,几乎所有的信息都会被存储在计算机上,这使得总体数据的获取成为可能。大数据思维不再采用传统统计的随机抽样模式,而是采用"样本即总体"的全数据思维模式。不再依赖于随机抽样,就可以分析更多数据,甚至是和某种现象相关的全体数据,从而可以更清楚地发现样本无法揭示的细节信息,为我们带来更全面的认知体验。

传统统计思维模式下进行数据处理时,以概率论为基础,根据样本特征推断总体特征。这种方法推断是否正确取决于样本的代表性。大数据思维强调的是使用全体数据。有了总体数据,我们就能清楚其实际分布的情况,而不再需要根据分布的假设来推断总体特征。和传统统计思维不同,大数据中的概率不再是事先设定,而是基于实际分布得出。

三、获取数据的方式不同

传统研究中以"定向型信息"为主,即人们通过设计调查表主动收集数据,逐个进行收集、整理。在回答一个特定的问题之前,人们会关心如何更好地收集数据,如实验设计和调查设计。因此传统数据的收集有很强的针对性,数据的提供者大多是确定的,身份特征是可识别的,还可以进行事后的核对。而在大数据时代以"发散型信息"为主,即对数据来源和产生者无过多的要求,亦非为了特定事物收集目的而产生。更重要的是,大数据时代以人工智能及物联网为背景,事物互联互通,数据实时产生,主动连接,定向汇集,且被人共享。大数据思维模式即基于此类多源数据进行分析寻找内在规律。例如,"淘宝"和"亚马逊"会将用户以往购买的物品和书的种类等数据主动连接和汇集起来,进行分析和判断用户需求,再将分析结果的信息反馈给老客户,方便老客户挑选自己喜爱的商品。这种智能"信息找人"的特点和功能用在现代医学上,可以使医疗科研工作者很方便地收集和处理疾病的信息,从而为疾病的预测、诊断和治疗提供帮助。

四、数据的性质不同

传统统计思维模式下,研究者难以容忍错误数据,非结构化数据要结构化后再分析。在传统统计工作中,无论在收集样本时,还是在做统计分析时,统计学家会用一整套的策略来控制偏倚、减少错误发生。在结果公布之前,也会检验样本是否存在潜在的偏倚。由于所收集的样本量小,因此有必要保证数据的结构化和精确化。换言之,传统统计数据具有样本量小,信息量丰富,针对性强,准确度高等性质。而大数据思维则不同,主要体现在以下两个方面。

一是高度容错机制。数据量越大,错误率越高,精度越低,即数据量往往与精度成反比,与错误率成正比。大数据的海量数据,不仅无针对性,而且垃圾信息多,错误多,但是错误的存在往往正是真实世界的一种体现。Google翻译系统是这方面较好的例证,尽管其输入源很混乱,但正是因为它可以接受有错误和混乱的数据,才使得它比其他翻译系统多利用成千上万的数据,从而使得其翻译质量越来越好[7]。大数据正是因为这种容错机制而大大提高了其预测的精度。

二是高度非结构化。大数据既包括文本数据,还包括图片、音频、视频、电子邮件、日志、地理位置以及聊天记录、支付记录等各种类别数据[8],这些数据结构混杂,格式不一。据估计,生物医学大数据中,仅有5%左右的数据是结构化数据,而95%的数据将是非结构化数据,其中蕴含着巨大的价值,有待挖掘。

五、分析方法的要求不同

在传统统计思维中,研究方法较为单一,主要依据统计方法,精确建模。统计模型基于一系列的假设,比如线性回归模型假设观测样本满足线性、独立性、正态性、方差齐性等条件。但是如果所提出的假设本身是不合理的,那么统计模型自然有偏,则无法反映事物局部细节特征和内在规律。依据目的提出假设后,再通过对收集的数据进行分析来验证其是否成立。因此,传统统计思维下的分析思路是"假设—验证"。

大数据以数据挖掘及智能算法为主要研究方法,快速、高效,且容错能力强。没有既定目标,没有理论模型,无需假设,而是通过特定的算法,对海量的数据进行分析,找出重要的特征和关系,从而发现其中隐藏的规律,然后进行判断和决策[9]。因此,大数据思维下的分析思路是"发现—总结"。亚马逊将发货"外包"给算法,让算法自动发货,正是

数据挖掘和智能算法的体现[10]。在医学上,由于人体的高度复杂性,理论上是难以精确建模的,此时基于医学大数据的数据挖掘及机器学习算法则可发挥其优势(如遗传算法,免疫算法,人工神经网络等)[11]。

六、医学领域的研究结论

在医学领域,由于传统统计方法可以通过验证假设大致提供疾病与危险因素之间的因果关系或验证某项干预措施的有效性,所以一般能以其为依据下结论。以随机对照试验(RCT)为例,其通过一系列统计方法的应用,可以提供高水平的试验来验证干预的有效性。毫无疑问,RCT 处于证据金字塔的顶端[12]。但是,也正是由于传统统计通常采用随机抽样方法获取样本,且存在严格的纳入排除标准,使 RCT 存在着局限性。怎样确保样本的代表性,减少错误,控制偏倚,使医学研究结论更为可靠,是统计学家和医学工作者必须思考的问题。

大数据分析可运用医学数据挖掘技术,对诸如基因变异和基因表达与各种疾病之间的相关性等问题进行研究,可对疾病的病因分析、诊断、治疗提供帮助。但是,由大数据分析得出的相关关系的结论并非具有因果性,在严谨的医学研究领域仅仅作为探索性研究的结论,或者为一些临床研究提供假设,其结论的可靠与否,还需要进一步的验证。

大数据思维把人们从传统的思维方式和价值观中解放出来,在公共交通、公共安全、社会管理,尤其是生物医学领域,均有巨大的应用价值。传统统计学思维是大数据科学发展的助推剂,而大数据亦是传统统计学科的发展契机。了解大数据思维和传统统计思维的差异,有助于取长补短,使二者有机融合。因此,生物统计教学应当在理解二者差异的基础上,与时俱进,使其适应生物医学大数据时代的发展要求,这是我们生物统计学工作者义不容辞的责任。在未来的教学工作中,应该顺应时代发展,引入大数据思维及数据挖掘技术,与传统统计教学的知识点互补融合,提升学生的专业素养及竞争力。

参考文献

[1] 宁康,陈挺. 生物医学大数据的现状与展望[J]. 科学通报,2015(5):534-546

[2] Bendler J,Wagner S,Brandt T,et al. Taming uncertainty in big data[J]. Business & Information Systems Engineering,2014,6(5):279-288

[3] 崔青云. 论统计思维及培养[J]. 山西煤炭管理干部学院学报,2009,22(3):34-35

[4] Mayer-Schnberger V,Cukier K. Big data:A revolution that will transform how we live,work and think[J]. Information,2014,17(1):181-183

[6] 冯启思. 数据统治世界[M]. 北京:中国人民大学出版社,2013:61-78

[5] 华键,杨文平,魏永越,等. 血生化指标与老年人帕金森病患病风险的相关性研究[J]. 中华老年医学杂志,2016,35(3):270-273

[7] Floridi L. Big data and their epistemological challenge[J]. Philos Technol,2012,25(4):435-437

[8] Hastie T. The elements of statistical learning:data mining,inference,and prediction[M]. 2nd ed. Springer,2009:21-45

[9] 张弛. 大数据思维范畴探究[J]. 华中科技大学学报(社会科学版),2015(2):120-125

[10] 徐子沛. 数据之巅:大数据革命,历史现实与未来[M]. 北京:中信出版社,2014:20-65

[11] 刘婵桢,王友俊. 医学数据挖掘技术与应用研究[J]. 生物医学工程学杂志,2014(5):1182-1186

[12] Zhang Z. Big data and clinical research:focusing on the area of critical care medicine in mainland China[J]. Quant Imaging Med Surg,2014,4(5):426-429