·中国医疗大数据·

医疗大数据的“欺骗性”及其对策

姜会珍[1]，马 琏[1]，朱卫国[1,2]

中国医学科学院 北京协和医学院 北京协和医院 [1]信息中心 [2]全科医学科（普通内科），北京 100730

通信作者：朱卫国 电话：010-69154149，E-mail：zhuwg@pumch.cn

【摘要】当前，针对医疗大数据的研究和应用越来越广泛，但毋庸置疑，医疗大数据本身具有一定欺骗性，在某些特殊场景下，可能会产生错误的结论和影响。本文从数据本身的欺骗性以及机器学习可能存在的陷阱展开，对医疗大数据产生欺骗性的原因进行分析；针对医疗大数据的欺骗性，从统计学角度阐述如何避免大数据陷阱；从模型角度分析模型被攻击的应对策略以及模型可解释性在医疗领域的重要性和方法。

【关键词】医疗大数据；欺骗性；机器学习

【中图分类号】R19-0；R195.4；C811 【文献标志码】A 【文章编号】1674-9081(2020)05-0542-05

DOI：10.3969/j.issn.1674-9081.2020.05.009

Medical Big Data “Deception” and Strategies

JIANG Hui-zhen[1], MA Lian[1], ZHU Wei-guo[1,2]

[1]Department of Information Center, [2]Department of Primary Care and Family Medicine (General Internal Medicine), Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China

Corresponding author: ZHU Wei-guo Tel: 86-10-69154149, E-mail: zhuwg@pumch.cn

【**Abstract**】At present, research and application of medical big data are more and more extensive. But inevitably, medical big data is of some deception, and in many scenarios, it can result in wrong conclusions and influence. In this paper, firstly we analyze the causes of medical big data deception from the data deception per se and pitfalls of machine learning. Then, we introduce how to avoid data pitfalls in statistics and analyze the strategies to tackle attacks on models. The importance and methods achieving model interpretability in the medical area are also mentioned.

【**Key words**】medical big data; deception; machine learning

Med J PUMCH, 2020,11(5):542-546

近年来，随着医疗信息研究水平的不断提升和医疗信息人才的多元化，针对医疗大数据的研究和智能模型的应用越来越广泛，甚至许多研究成果已开始应用于临床[1]，在减轻医务/管理人员工作负担的同时，亦有助于减少医院不良事件发生，为患者提供更精准、有效的诊疗服务[2]。医疗领域科学、严谨的特性决定人们对医疗大数据的准确性和可靠性具有非常严苛的要求，但大数据本身具有一定的欺骗性。Chan 等[3]在对精神疾病患者的生物标志物研究中发现，研究结果再现性差的主要原因是欺诈、不恰当的

基金项目：国家重点研发计划（2018YFC0116905）；中国医学科学院医学与健康科技创新工程（2016-I2M-2-004）；美国中华医学基金会公开竞标项目（CMB-OC）（16-258）

利益冲突：无

统计分析等。Ranstam 等[4]研究发现，医学研究中欺诈行为如伪造、篡改数据，欺骗性设计、分析等均为不可忽视的行为。除了大众所熟知的“系统误差”，还有数据陷阱以及因模型的脆弱性所带来的风险。Goodfellow 等[5]认为，对于机器学习模型，数据集中一些小的干扰可能导致模型输出错误的结果。如何发现医疗大数据挖掘分析中的陷阱，并采取相应的策略来减少医疗大数据的欺骗性至关重要。本文对医疗大数据的欺骗性原因进行梳理和总结，并从统计学角度阐述如何避免大数据陷阱，从模型角度分析模型被攻击的应对策略以及模型可解释性在医疗领域的重要性和方法。

1　医疗大数据的欺骗性相关概念

医疗大数据的欺骗性是指在医疗大数据研究中，因被动或主动干预造成研究结果不正确的现象。本文主要从数据的欺骗性和机器学习陷阱两个方面概述。数据的欺骗性是指用于医疗大数据研究的样本数据在选取或处理时，由于处理不当而造成的偏差等；机器学习陷阱是指在医疗大数据的训练过程中，因模型问题导致结果不准确或被攻击。图 1 为医疗大数据研究基本方案及流程[6]，数据的欺骗性和机器学习陷阱分别对应图中①和②常见隐患，同时，步骤①分析结果也将直接影响特征工程效果。因此，对于医疗大数据相关研究来说，数据的欺骗性和机器学习陷阱在整个建模过程中均应尽量避免，以提高模型预测结果的可信度。

1.1　数据的欺骗性

由于数据在结论展现前需经过取样、清洗、建模、分析以及应用等过程。Dallachiesa 等[7]提出通过数据清洗系统来减少“脏数据”，保障数据质量。Rahm 等[8]认为，数据处理工作对提高数据质量至关重要，并且其阐述了数据清洗、处理的方法。即使通过清洗等方法清除部分异常数据，从统计学角度来看，大数据仍具有欺骗性，主要分为选择偏倚、结果的局限性和数据噪声。

1.1.1　选择偏倚

有一种错误认知是大数据至上，但实际上，数据集本身和数据分析并非完全客观，在大数据采集和分析中会存在各种偏差。若过分相信大数据总能反映、揭示真理，则称为“大数据自大”。Pauleen 等[9]提出应合理管理和使用大数据，若过度使用/滥用，将会导致一系列问题如金融危机。典型的几类造成数据偏差的原因包括：第一，选择误差。如果选择的数据样本分布不均匀即会出现选择误差。例如，在机场做问卷调查，期望对全民健康水平进行评估，则注定是失败的，因为机场人群的分布和全国人群分布不一致，不具有代表性，样本选择具有偏差。第二，幸存者误差。若有些样本数据无法采集即会出现幸存者误差。例如，为评估某药物对患者的副作用，选取存活患者展开调查，因无法获取药物试验中已故患者的数据，而这些患者可能是发生药物副作用较多的人群。因此这样的采样并不全面，将导致分析结果不正确。第三，数据真实性存疑。在研究中，参与者因个人利益等原因可能会出现一些欺骗行为，这会降低研究数据的质量[10]。因此，应尽可能增大研究的数据量，减小错误数据对研究结果的干扰。

1.1.2　结果的局限性

结果的局限性是引起数据欺骗性的常见原因。无论是数据统计分析，还是训练机器学习模型，均是在有限数据中进行局部归纳推理，并泛化至全局样本空间中。可用如下公式来表示：$Y=F(X)$。

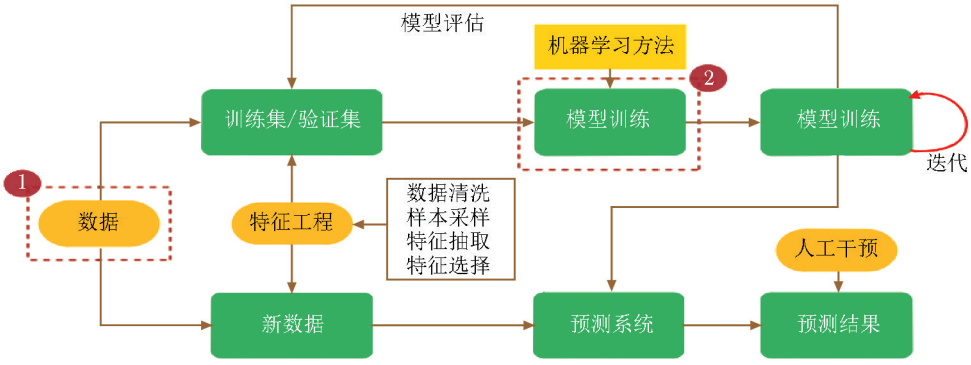

图 1　医疗大数据研究过程

该过程可被描述为学习一个目标函数 F，F 能最好地将输入变量 X 映射至输出变量 Y。其本质是试图通过找到的变量相关性去论证因果关系。但由于因果变量相关性存在多种可能性，理论上来说，只要有超大样本和多个变量进行足够多次的建模，均可能找到各种看似合理的相关性，其完全符合统计方法，但采用这样的相关性来论证因果关系具有不可信性。比如，研究肿瘤患者入院等待时间与预后的关系，数据分析表明入院等待时间越长，患者预后越好；反之，预后越差。而实际原因是紧急入院患者通常病情更重，因而预后相对较差。患者入院等待时间与其预后本无关联，但在数据上却表现为相对一致。Rohrer[11]研究提出，数据具有相关性并不意味着有因果关系。如何判断数据之间的关系是否为真正的因果关系呢？Simon[12]提出通过引入其他变量、公式或参数来检验数据之间的相关性是否真实。

1.1.3 数据噪声

噪声数据是指存在错误或异常（偏离期望值）的数据，这些数据能干扰分析结果。在将统计学应用于大数据分析时，应提防数据噪声以及数据背后逻辑和动机不透明所带来的风险。2008 年，谷歌（Google）公司领衔在 *Nature* 上发表论文，推出“谷歌流感趋势”（Google Flu Trends）预测。其根据互联网上有关流行性感冒的搜索数量和分布来估计各地区流行性感冒类疾病的患者数目，开发了具有较高准确性和实时性的预测系统[13-14]。但 2013 年 Butler[15]指出，“谷歌流感趋势”在 2012 年的预测结果比实际数据高了 1 倍多。经分析，是由于媒体对此段时间的美国流行性感冒类疾病作了渲染，使许多非流行性感冒患者也进行了相关搜索，从而干扰了“谷歌流感趋势”的预测。在统计学中，这被称为系统误差，样本数据量再大也无法避免。

1.2 机器学习陷阱

除了数据的欺骗性，在建模过程中也存在机器学习陷阱，导致试验结果存在一定偏差，包括模型本身的缺陷、模型选择不当和模型对抗性攻击。

1.2.1 模型本身的缺陷

“黑天鹅”理论[16]在大数据领域是热门课题，其蕴含的逻辑是未知的小概率事件，一般无法预测，而其一旦发生将会产生巨大的影响。归纳和演绎是大数据挖掘常用的两个基本手段，前者是从具体的事件中归纳出一般性规律，即从特殊到一般的泛化过程；后者是从基础原理推演出具体的情况，即从一般到特殊的特化过程。大数据挖掘通常从有限的数据中进行局部归纳推理，并将结论推广到全局样本空间中。但这样的归纳推理不仅脆弱且蕴含一定风险。近年来，基于日积月累的个性化医疗信息数据，越来越多的研究开始致力于疾病的诊断预测，如 Siuly 等[17]提出计算机辅助诊断系统在神经系统疾病诊断方面的应用。但这样的疾病预测模型很难预测到未知的新疾病，如严重急性呼吸综合征（severe acute respiratory syndrome，SARS）、甲型 H1N1 流感、埃博拉病毒的暴发等“黑天鹅”事件。因此，模型认为小概率事件不会发生，显然这样的假设会导致完全依赖于大数据的决策存在风险[18]。

1.2.2 模型选择不当

在需要用机器学习来解决医疗大数据中的具体问题时，模型选择至关重要。随着机器学习理论和技术的快速发展，已有足够多的模型可作为解决问题的工具。按照主流的分类方法，其包括监督学习、无监督学习、半监督学习、强化学习、主动学习等，有监督学习可细分为线性模型、树模型、深度模型等。实际应用时，需根据数据的形态、问题的类型、期望达到的目标来选择适合的模型。如果面对的问题不太明确或数据形态不常见，缺乏经验的建模师在建模时很容易出现偏差，造成模型性能较差，无法达到预期。例如，医疗临床数据包含不同值域的数值变量、类别变量以及布尔变量，其比较适合用树模型或深度模型，而非线性模型。另外，Doornik 等[19]研究显示，模型选择不当易产生一些虚假的数据关联，且其阐述了如何进行模型选择。

1.2.3 模型对抗性攻击

像软件系统有安全漏洞一样，机器学习模型也存在漏洞，甚至更脆弱，在受到外部恶意攻击时模型决策被干扰。“谷歌大脑”在 2018 年的研究表明[20]，任何机器学习模型均可以被欺骗、攻击，从而得出不正确的预测结果，且攻击者几乎可以让模型输出任何想要的结果。大部分模型攻击方式是对抗性攻击，即在正常样本中加入一定的扰动来干扰模型。机器学习模型由一系列特定的参数计算和变量变换组成，这种变换对输入的微小变化非常敏感，利用这种敏感性来修改甚至是控制模型是攻击者常用的手段。这是人工智能安全领域中一个重要的课题，特别是在医疗大数据领域，人们对机器学习的临床应用一直持有谨慎保守的态度。保证模型的稳健性、避免其被攻击尤其重要。

2 医疗大数据欺骗性应对策略探讨

医疗大数据的欺骗性应对策略可从数据和模型两个角度进行概述。

2.1 避免数据欺骗

2.1.1 确保取样的代表性

从医疗大数据研究的流程上来看，首先应确保样本选取具有代表性。理论上来讲，大数据的特点之一是研究全体，而非抽样数据，但在实际研究中很难获得全部数据，而是需要基于能获得的数据进行分析。数据的欺骗性多与此有关，数据的样本选取代表性差是制约模型性能的根本因素之一。依据机器学习的原始假设，高质量的训练样本应最接近真实样本分布。因此，为了让模型达到最佳效果，在数据采样时应保证采样候选集的数据分布与真实样本分布一致或尽可能接近。同时，采样方法应保证客观且随机，以避免人为主观因素导致的数据倾向。

2.1.2 尊重客观逻辑

在规范数据样本选取后，对数据进行探索性分析应注意尊重数据的客观逻辑，保证数据分析的合理性。经验欠缺的建模师在挖掘分析数据之间的规律时，往往会根据个人经验假定两个变量之间存在某种关联，然后通过数据分析或模型去验证。有时为了达到预期的结果，会给两个无关变量强行建立某种关联。因此，应尊重数据的客观逻辑，避免强行加入个人主观因素，如前文患者入院等待时间与预后的关系分析案例。

2.1.3 基于数据演化更新分析模型

经过规范的数据样本选取和数据分析后，需注意如有数据演化情况应及时更新模型。数据是模型的根基，数据的演化可能会产生一些数据噪声甚至使数据分布偏离训练集原本的形态，对模型的预测性能产生极大影响。因此，在建模时需考虑数据未来的演化情况，提前作出判断并修正方案。通常来说，存在数据演化的场景模型需定期重新训练并更新。

2.2 防御模型被对抗性攻击

2.2.1 对抗样本检测

对抗样本即用于攻击模型的不良数据，该部分数据不属于正常样本数据，目的是干扰模型的正常训练或预测。对抗样本检测是指在模型训练或预测前构造一个对抗样本检测器，对正常样本和对抗样本加以区分，并作相应处理。Feinman 等[21]提出，通过深度神经网络可有效区分对抗样本和正常样本，经受试者工作特征曲线验证其曲线下面积可达 0.8~0.93。

2.2.2 还原对抗样本

对抗样本一般是人为对原始样本处理后的数据。对于对抗样本，可通过对抗样本检测器加以识别，同时将对抗样本还原为初始样本，保障数据无误。

2.2.3 增强模型

增加样本量以保证模型训练的稳健性。模型稳健性越好，对抗样本对其产生的干扰越小。应用较多的方案是收集或构造更多的样本，甚至将对抗样本加入模型训练，同时在模型中加入正则项以防止模型过拟合，即防止其训练数据过于敏感，从而保证模型的稳健性。

2.3 保证模型可解释性

对于机器学习模型，线性模型具有可解释性，而非单棵的树模型和深度学习模型不具有可解释性。Lipton[22]阐述了可解释性模型的特点，并对不同模型的可解释性作了对比分析。Poursabzi-Sangdeh 等[23]通过对照试验评估特征的数量和模型的透明度（是否为黑盒子）对模型可解释性的影响。医疗大数据不同于其他行业，用于医疗大数据研究的机器学习模型需具有更强的可解释性，以确保医疗安全。因此，在进行医疗大数据相关研究和应用时，应尽可能保证模型的可解释性：（1）特征主导模型预测。尽量找出在实际场景中特征的相互作用，以了解在建模过程中如何建设特征工程。（2）模型可验证。可通过曲线下面积、精确度等指标评估模型有效性，保证每一个特征的有效性均可被充分验证。

3 总结与展望

医疗大数据分析在提供精准、有效诊疗服务的同时，其也具有欺骗性。本文从数据的欺骗性和机器学习陷阱两方面介绍了医疗大数据欺骗性的原因及分类，并从统计学角度和模型角度分析应对策略，以减少医疗大数据研究过程中可能造成的差错。医疗领域严谨的特性决定了其对数据的准确性、模型决策合理性要求极其严格，但现阶段针对医疗大数据的欺骗性以及应对策略的研究尚缺乏深度，尤其针对模型对抗性攻击方面的应对策略尚需深入研究，以保障医疗大数据应用的安全性。

作者贡献：姜会珍、朱卫国提供论文思路；姜会珍撰写论文；朱卫国、马琏修改论文。

参 考 文 献

[1] Lee CH, Yoon HJ. Medical big data: promise and challenges [J]. Kidney Res Clin Pract, 2017, 36: 3-11.

[2] Price WN, Cohen IG. Privacy in the age of medical big data [J]. Nat Med, 2019, 25: 37-43.

[3] Chan MK, Cooper JD, Bahn S. Commercialisation of biomarker tests for mental illnesses: advances and obstacles [J]. Trends Biotechnol, 2015, 33: 712-723.

[4] Ranstam J, Buyse M, George SL, et al. Fraud in medical research: an international survey of biostatisticians [J]. Controll Clini Trials, 2000, 21: 415-427.

[5] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint, 2014, 1412. 6572.

[6] Erickson BJ, Korfiatis P, Akkus Z, et al. Machine learning for medical imaging [J]. Radiographics, 2017, 37: 505-515.

[7] Dallachiesa M, Ebaid A, Eldawy A, et al. NADEEF: a commodity data cleaning system [C] //Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, ACM, 2013: 541-552.

[8] Rahm E, Do HH. Data cleaning: Problems and current approaches [J]. IEEE Data Eng, Bull, 2000, 23: 3-13.

[9] Pauleen DJ, Rooney D, Intezari A. Big data, little wisdom: trouble brewing? Ethical implications for the information systems discipline [J]. Soc Epistemol, 2017, 31: 400-416.

[10] McCaul ME, Wand GS. Detecting deception in our research participants: are your participants who you think they are? [J]. Alcoholism Clin Exp Res, 2018, 42: 230-237.

[11] Rohrer JM. Thinking clearly about correlations and causation: Graphical causal models for observational data [J]. Advances in Methods and Practices in Psychological Science (AMPPS), 2018, 1: 27-42.

[12] Simon HA. Spurious correlation: A causal interpretation [J]. J Am Stat Assoc, 1954, 49: 467-479.

[13] Wilson N, Mason K, Tobias M, et al. Interpreting "Google Flu Trends" data for pandemic H1N1 influenza: the New Zealand experience [J]. Euro Surveill, 2009, 14: 19386.

[14] Lazer D, Kennedy R, King G, et al. Big data. The parable of Google Flu: traps in big data analysis [J]. Science, 2014, 343: 1203-1205.

[15] Butler D. When Google got flu wrong [J]. Nature, 2013, 494: 155.

[16] Taleb NN. The black swan: the impact of the highly improbable [M]. New York: Random house, 2007.

[17] Siuly S, Zhang Y. Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis [J]. Data Science and Engineering (DSE), 2016, 1: 54-64.

[18] Batrouni M, Bertaux A, Nicolle C. Scenario analysis, from BigData to black swan [J]. Comput Sci Rev, 2018, 28: 131-139.

[19] Doornik JA, Hendry DF. Statistical model selection with "Big Data" [J]. Cogent Economics & Finance, 2015, 3: 1045216.

[20] Elsayed G, Shankar S, Cheung B, et al. Adversarial examples that fool both computer vision and time-limited humans [C] //Advances in Neural Information Processing Systems, 2018: 3910-3920.

[21] Feinman R, Curtin RR, Shintre S, et al. Detecting Adversarial Samples from Artifacts [J]. arXiv preprint, 2017, 1703. 00410.

[22] Lipton ZC. The mythos of model interpretability [J]. Queue, 2018, 16: 31-57.

[23] Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, et al. Manipulating and Measuring Model Interpretability [J]. arXiv preprint, 2018, 1802. 07810.

(收稿日期: 2019-07-15)