

基于大数据的卫生统计工作开展方法研究

刘荣成

昭通市妇幼保健院, 云南昭通, 657000;

摘要: 大数据技术的快速发展为卫生统计工作提供了新的机遇, 传统的统计方式在数据采集、分析、存储等方面存在局限, 而大数据的引入能够极大提升数据处理的效率和准确性。当前, 医疗行业的数据量呈爆发式增长, 如何合理利用这些数据, 使其在公共卫生决策、疾病预防、资源配置等方面发挥更大作用, 成为卫生统计工作者面临的重要课题。本文通过深入研究分析了大数据下的卫生统计工作开展方法, 以期卫生统计工作提供更科学的思路和方法。

关键词: 大数据; 卫生统计工作; 应用场景; 开展方法

DOI: 10. 64216/3080-1486. 25. 06. 034

卫生领域产生的数据呈现出爆炸式增长的趋势, 包括电子病历、医疗影像、生物样本、可穿戴设备等多种类型。这些数据蕴含着丰富的健康信息, 如果能够有效地加以利用, 必将极大地推动卫生事业的发展。然而, 面对如此庞大而复杂的数据, 传统的统计方法已经难以胜任。我们亟需引入大数据技术, 建立一套科学、高效、实用的卫生统计工作流程, 才能充分发挥卫生统计工作的效果。

1 卫生统计工作开展中大数据的应用场景

1.1 公共卫生政策评估

如今, 借助电子病历、医保结算数据、居民健康档案以及互联网健康平台的数据, 政策制定者可以分析政策实施前后的健康指标变化。例如, 某地实施儿童疫苗接种补贴政策后, 政府可以利用大数据追踪疫苗接种率的提升情况, 同时结合医院诊疗数据分析儿童传染病发病率是否下降^[1]。如果发现政策实施后接种率提升但疾病发病率变化不大, 可能意味着疫苗效果有限或其他防控措施尚未完善。此外, 大数据还能用于分析医疗费用控制政策的成效, 比如医保控费措施是否真正减少了不必要的医疗支出, 或者是否影响了患者获得必要治疗的机会。卫生部门不仅能通过这些数据评估政策的有效性, 还能及时调整方案, 使其更符合实际需求。

1.2 医疗资源配置优化

医疗资源分布的不均衡一直是医疗体系面临的挑战, 尤其是在偏远地区, 优质医疗资源相对匮乏, 而城市大医院则面临就诊高峰和资源浪费并存的局面。大数据分析能够帮助决策者更合理地配置医疗资源, 提高整体医疗体系的运作效率。例如, 通过分析不同地区的就诊数据、住院率、急诊需求以及疾病负担情况, 可以预

测某些地区对特定医疗资源的需求量, 为医院建设、医疗设备投放和医生调配提供科学依据。此外, 急诊科往往面临资源紧张的问题, 利用大数据分析患者流量、病种分布和高峰时段, 可以优化急诊分诊流程, 减少患者等待时间, 提高急救效率。

1.3 健康管理服务创新

传统的健康管理主要依赖个人的主观描述和有限的体检数据, 而大数据技术可以整合个人健康档案、可穿戴设备监测数据、生活习惯记录等信息, 形成更全面的健康画像。例如, 智能手环和智能手表可以实时监测心率、血压、睡眠质量等数据, 结合长期健康数据分析, 智能健康管理系统能够识别个体的健康风险, 并向用户提供个性化的健康建议。如果一个用户的心率异常, 系统可以提醒他及时就医, 甚至在必要时向医院或家属发送警报。对于慢性病管理, 大数据能够帮助医生更精准地制定个性化治疗方案, 例如糖尿病患者的血糖监测数据可以与饮食、运动等因素结合分析, 医生可以据此调整用药方案, 提高疾病管理的精准度。

2 基于大数据的卫生统计工作开展方法

2.1 卫生统计大数据的采集方法

2.1.1 传统数据采集方式的优化

优化的关键在于减少人为干预, 提高数据自动化采集水平, 同时加强对数据填报人员的培训和审核机制。电子化报送系统的升级, 使数据录入更加便捷, 同时结合智能校验技术, 能够有效降低人为错误。医院、疾控中心等机构的数据采集系统逐步向电子病历、区域卫生信息平台等方向集中, 减少了纸质报表的依赖, 提高了数据的实时性和可追溯性。此外, 基于移动端的填报工具也在推广, 使得基层医疗机构和社区卫生服务人员能

够更方便地提交数据,缩短了数据汇总周期。

2.1.2 新兴数据采集技术的引入

物联网技术的应用,使得可穿戴设备、智能医疗设备的数据能够直接接入统计系统,减少了人工录入环节,提升了数据的实时性和准确性。大数据采集技术的发展,使得社交媒体、互联网搜索、在线问诊等非传统数据源逐步被纳入卫生统计分析范围,为疾病监测、公共卫生决策提供了更丰富的信息支撑。自然语言处理和图像识别等人工智能技术,使非结构化数据的利用率大幅提升,医生的电子病历、影像资料等数据能够被自动解析并转化为可统计的结构化数据,提高了数据整合的效率。

2.1.3 多源异构数据的整合策略

医院信息系统、疾控机构监测数据、政府健康档案、互联网健康数据等数据源在格式、标准、存储方式上存在较大差异,统一数据标准和建立高效的数据融合机制是解决问题的关键。建立基于大数据平台的数据共享体系,使各类数据能够按照统一的标准进行存储和管理,提升数据的兼容性和可用性^[2]。数据清洗和标准化技术的应用,使得不同格式的数据能够在同一平台上进行分析,避免因数据来源不同而导致的信息孤岛问题。隐私保护和数据安全同样是多源数据整合过程中需要重点考虑的问题,采用数据脱敏、分布式存储和访问权限控制等技术,确保数据在共享过程中不会泄露个人隐私。

2.2 卫生统计大数据的预处理技术

2.2.1 数据清洗方法

数据清洗是整个预处理流程的基础,主要包括去除重复数据、修正错误数据、统一数据格式等。卫生统计数据往往来源于多个信息系统,不同机构的数据录入格式可能存在差异,甚至同一数据源内部也可能出现冗余记录。如果不进行去重和格式调整,后续的分析结果可能会受到影响。通常,去重方法包括哈希去重、主键匹配或者基于相似度的去重算法,而错误数据的修正则需要结合业务规则,例如年龄字段中出现负数或超过合理范围的值,就需要进行修正或删除。

2.2.2 数据标准化策略

卫生统计数据通常涉及多种单位、编码和分类方式,例如血压单位有 mmHg 和 kPa 两种表示方式,疾病编码可能采用 ICD-9 或 ICD-10 标准,如果不进行转换,就可能导致数据无法兼容,影响分析的准确性。标准化的方法一般包括单位换算、编码转换和分类重映射,不仅要考虑数据本身的转换规则,还要确保符合国家或行业制定的标准。例如,医院上报的病种数据可能使用本地编码,在数据整合时需要转换为国际通用编码,以便于跨机构和跨地区分析。

2.2.3 缺失值处理方案

缺失值的处理是数据预处理中不可忽视的环节,直

接删除可能会导致样本量减少,影响统计分析的代表性,而不处理则会影响模型的稳定性。处理缺失值的方法主要包括删除、填充和插值。删除适用于缺失比例较小且随机分布的情况,如果某个字段的缺失率过高,删除这一特征可能比填充更合理。填充的方法较为常见,例如用均值、中位数或众数代替缺失值,或者在分类变量中用“未知”作为占位符,以保持数据完整性。插值则适用于时间序列数据,例如患者的体温记录如果某个时间点缺失,可以用前后温度的平均值填补,以保持数据的连续性。

2.2.4 异常值识别技术

异常值识别在卫生统计数据处理过程中尤为关键,异常值可能来源于录入错误、设备故障或者个体差异,如果不加以处理,可能会影响整体分析结果。常见的识别方法包括基于统计学的方法、机器学习算法和业务规则设定。例如,基于均值和标准差的 3σ 原则可以识别大部分异常值,即某个数据点如果超出均值 ± 3 倍标准差的范围,就可能是异常数据^[3]。此外,箱线图法可以通过四分位距(IQR)设定阈值,识别超出正常范围的值。如果某项指标的异常值符合医学常识,比如某些疾病可能导致异常高血糖或低血压,就不能简单剔除,而是需要结合业务逻辑进行分析。

2.3 卫生统计大数据的存储管理

2.3.1 分布式存储架构设计

面对海量数据,传统的单机存储模式早已无法满足需求,分布式存储架构成为主流选择。分布式存储架构的设计需要考虑数据的高可用性、扩展性以及容灾能力。常见的方式是采用分片存储和副本机制,将数据分散存放在不同的节点上,避免单点故障带来的数据丢失风险。同时,合理设计数据分布策略,使数据读写均衡,提升系统整体性能。主流的分布式存储方案包括 HDFS、Ceph 等,它们支持大规模数据的高效存储,并且具备良好的扩展能力,适用于卫生统计大数据的存储需求。

2.3.2 数据库选型及优化

面对卫生统计数据的复杂性,关系型数据库和非关系型数据库需要结合使用。对于结构化数据,如人口统计、医疗资源配置等信息,传统的 MySQL、PostgreSQL 等关系型数据库仍然是可靠的选择,通过索引优化、分区表等方式提升查询效率。而对于非结构化数据,如医疗影像、诊疗记录等,NoSQL 数据库如 MongoDB、Cassandra 或分布式 KV 存储数据库如 HBase 更具优势。混合数据库架构能够兼顾不同类型数据的存储需求,提升系统灵活性。此外,数据库优化也是确保高效存储管理的重要环节,包括合理的索引设计、缓存机制的应用以及分布式事务的优化,避免数据处理过程中出现性能瓶颈。

2.3.3 数据安全防护措施

一方面,访问控制是基础,采用基于角色的权限管理(RBAC),确保不同用户只能访问其权限范围内的数据,防止数据滥用或泄露。另一方面,日志审计机制能够有效监控数据访问情况,及时发现异常操作,提升系统的安全性。为了应对可能的数据丢失风险,定期备份和异地容灾机制是必不可少的,采用多副本存储策略,确保即使发生硬件故障或灾难性事故,数据仍然能够快速恢复,保障卫生统计数据的持续可用性。

2.4 卫生统计大数据的分析挖掘

2.4.1 描述性统计分析方法

卫生统计大数据的分析挖掘涉及多个层面的统计方法和技术手段,首先是描述性统计分析,这一部分主要用于整理和总结数据的基本特征,帮助研究人员快速了解总体情况。常见的方法包括均值、中位数、标准差、频数分布等,这些指标能够直观地反映卫生数据的集中趋势和离散程度。例如,在流行病学研究中,统计某种疾病在不同年龄段人群中的患病率,或者分析不同地区居民的健康状况差异,都是描述性统计的典型应用^[4]。面对大规模数据,仅仅依赖传统的统计方法难以充分展现数据特征,因此,探索性数据分析(EDA)在卫生统计中也越来越重要,它强调数据的可视化和模式识别,能够帮助研究人员发现潜在的趋势和异常值,从而为后续的深入分析提供方向。

2.4.2 推断性统计分析方法

在描述性统计的基础上,推断性统计方法用于从样本数据推测总体特征,以支持决策和政策制定。假设检验、回归分析和方差分析是常见的推断性统计方法。以医院的疾病诊疗数据为例,如果想判断某种治疗方案是否显著降低疾病复发率,研究人员通常会采用假设检验的方法,通过统计检验结果来推断治疗方案的有效性。此外,回归分析被广泛用于研究多个变量之间的关系,比如分析环境污染对居民健康的影响,或者探讨生活习惯与慢性病发病率之间的联系。在大数据背景下,传统的统计推断方法逐渐与更复杂的计算手段相结合,例如贝叶斯分析和生存分析等,进一步提高了数据分析的精准度,为卫生领域的科学研究和决策提供了有力支持。

2.4.3 机器学习算法的应用

随着数据规模的扩大和复杂性的增加,传统统计方法在处理非线性关系、海量数据和高维度变量时存在一定的局限性,因此,机器学习算法在卫生统计数据分析中的应用日益广泛。监督学习、无监督学习和强化学习等不同类型的算法在实际问题中各有应用场景。例如,

决策树、随机森林和支持向量机等监督学习算法可以用于疾病预测和风险评估,通过训练大量历史病例数据,建立精准的预测模型,帮助医生在临床诊断时做出更科学的决策。无监督学习方法,如聚类分析和主成分分析,则适用于模式识别和数据降维,比如在健康管理领域,利用聚类算法对人群健康状况进行分类,制定个性化的健康干预策略。

2.4.4 可视化分析技术

数据分析不仅仅是计算和建模,如何让分析结果更直观、易于理解同样至关重要。可视化分析技术在卫生统计大数据挖掘中发挥着重要作用,一方面可以用于探索性数据分析,帮助发现数据中的异常点和趋势,另一方面也能提高数据呈现的直观性,使决策者和研究人员更容易解读分析结果。传统的条形图、折线图、散点图等仍然是常用的可视化工具,而随着数据分析技术的进步,交互式可视化和地理信息系统(GIS)等技术也逐渐应用于卫生统计领域^[5]。例如,在疾病传播研究中,利用GIS技术绘制疫情地图,可以动态展示疾病在不同地区的传播情况,为公共卫生决策提供直观依据。

3 结语

总的来说,大数据技术的应用,使卫生统计工作从传统的静态分析向动态、智能化分析转变,为医疗卫生管理提供了更高效的决策支持。但与此同时,数据质量、隐私保护、技术人才等问题也需要得到充分重视。未来,卫生统计工作应在技术创新、管理优化、人才培养等方面持续探索,不断提升数据的价值挖掘能力,让卫生统计真正发挥出支撑医疗卫生事业发展的重要作用。

参考文献

- [1]于石成,肖革新,郭莹.大数据视角下的卫生统计工作[J].医学信息学杂志,2013,34(10):47-50.
- [2]任奕先.探讨大数据下卫生统计工作的开展[J].通讯世界,2016,22(5):27-28.
- [3]谢红,周明,党少农,曾令霞,李强,康轶君.数据库在卫生统计工作中的应用[J].中国医学教育技术,2002,16(5):314-316.
- [4]董宏.大数据背景下开展乡镇统计工作[J].财富生活,2024(21):78-80.
- [5]曾玉琴,刘刚琴.新时期加强医院卫生统计工作的思考[J].科学与信息化,2024(9):141-143.

作者简介:刘荣成(1991-01),男,汉族,云南昭通人,本科,统计师,主要研究方向:卫生健康统计。