
一些人工智能疾病预测模型基于可疑数据训练而成

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/39318.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

一些人工智能疾病预测模型基于可疑数据训练而成

。近日，一篇公布于预印本服务器medRxiv的论文显示，一些可疑的数据集被用于训练旨在预测人们中风和糖尿病风险的人工智能（AI）模型。据《自然》报道，一些模型似乎已临床应用，目前尚不清楚这是否导致了错误的诊断结果。目前，至少有两家期刊正在调查使用这些数据集的研究。

图片来源：Marko Nikolic/Alamy

澳大利亚昆士兰科技大学的统计学家Adrian Barnett团队查阅了 124 篇经过同行评审的论文，这些论文均报告使用了两个开放获取的健康数据集中的一种以上来训练机器学习模型，但关于数据来源的信息却寥寥无几。

新研究揭示了多个在真实人群数据中不应出现的异常之处，这让Barnett团队怀疑数据可能是伪造的。“这种情况实在令人大吃一惊。”Barnett表示。

“基于来源不明数据训练的预测模型在临床决策中毫无用武之地。它们本质上就是不可靠的。”澳大利亚乔治全球健康研究院的公共卫生研究员Soumyadeep Bhaumik表示，如果这些工具未使用真实世界数据，就可能做出错误预测，导致临床医生作出不当决策，例如不必要地开具治疗方案，或在需要时未予开具。

据悉，至少有两个模型已被印度尼西亚和西班牙的医院采用，其中一个还出现在2024年提交的一份医疗器械专利申请中。

Bhaumik表示，研究机构和资助方必须要求研究人员披露用于训练医疗应用AI模型的数据来源，而期刊应拒收未披露数据来源的论文。Barnett提醒，研究中标记出的数据集现在应被“下架”，以防后续研究继续使用。

该研究所涉及的两个数据集来自Kaggle——
一个供开发者获取数据集以构建机器学习模型的平台。

第一个数据集名为“中风预测数据集”，上传时的描述为“用于预测中风事件的11项临床特征”。该数据集包含5110人的健康信息，包括心脏病史、婚姻状况、平均血糖水平和身体质量指数（BMI）等风险因素数据。

然而，当研究人员将平均血糖水平与参与者标识符进行关联分析时，却发现了若干异常情况。Barnett表示，其中一个异常是数据点缺失极少，这与真实数据形成鲜明对比——真实数据往往存在缺失，因为部分参与者会错过随访、退出研究或去世。“现实世界中，任何数据集都不可能完全完整。”他表示。

共有104篇研究论文曾使用该数据集构建中风预测模型，其中包括印度尼西亚某医院采用的模型，以及在少数人身上进行测试的模型。

该中风数据集由西班牙数据科学家Federico Soriano Palacios上传，量已超过28.8万次。在Kaggle平台关于该数据集的讨论区中，Palacios声明数据源自保密渠道，仅限用于教育目的。

第二个数据集名为“糖尿病预测数据集”，被描述为“一个利用医疗和人口统计数据预测糖尿病的综合数据集”。该数据集包含10万人的信息，包括BMI、吸烟史和血糖水平。但Barnett团队发现，所有参与者的血糖数据仅包含18个离散值。鉴于人群间存在的巨大差异，这种情况是不可能的。该团队还表示，他们发现了数千个似乎重复的数值。

有21项研究使用了这一数据集来构建糖尿病预测模型，但截至目前，这些模型均未在临床应用。值得一提的是，有一项研究同时使用了这两个数据集。

糖尿病数据集由印度数据工程师Mohammed Mustafa上传，他在Kaggle上表示，这些数据来源于汇

总的电子健康记录。针对讨论区中一位用户的提问，Mustafa表示：“出于保密原因或其他限制，我无法透露该糖尿病预测数据集的具体来源。”

截至目前，Palacio和Mustafa仍未对数据来源作出回应，Kaggle也拒绝就该平台是否会对这些数据集进行调查或采取任何行动发表评论。

上个月，研究人员已在PubPeer网站上发表了关于这124篇论文的担忧。PubPeer是一个供科学家和研究诚信调查人员讨论已发表论文的网站。其中一些论文的作者已作出回应。

印度SRM科技学院的计算机工程师M. Karpagam是2025年4月发表于《科学报告》的一篇论文的第一作者。论文报告研究人员利用这两个数据集开发出一种模型，用于预测残疾人的糖尿病和中风严重程度。Karpagam在PubPeer回应称，该数据集仅用于验证AI框架的性能，而非得出临床结论。在另一条评论中，她写道：“在开展和提交该研究时，我们并不知晓有任何报告表明这些数据集有可能是合成或模拟的。”她强调，在机器学习研究领域，来自“公开可用的数据库”的数据集通常被用作开发和评估预测模型的基准数据集。

《科学报告》主编Rafal Marszalek表示，其团队正在审查中Barnett团队论文中标记的5篇论文——包括Karpagam的论文，并将视情况采取相应措施。自3月31日以来，该期刊已撤回其中3篇论文，并指出“这些研究中使用的数据来源和有效性存疑”。

《生物工程与生物技术前沿》也在PubPeer上表示，正在评估Barnett团队标记的另一篇论文中相关问题，该期刊的研究诚信团队将“根据出版伦理委员会的指导方针对相关问题进行彻查”。

相关论文信息：<https://doi.org/10.64898/2026.02.24.26347028>

作者：文乐乐 来源：中国科学报

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发