医疗AI模型存在隐私风险

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/40551.html

医疗AI模型存在隐私风险。一项研究发现，那些贡献自身数据用于医疗人工智能（AI）模型训练的个人，可能面临在网络攻击中被识别的风险。此外，代表性不足群体面临的数据泄露风险也可能更高。研究人员指出，当前的风险评估并未将这些群体纳入考量，他们呼吁采取进一步的风险缓解措施并实施严格的访问控制。相关研究成果6月24日发表于《自然》。

医疗AI模型有望改善全球健康状况，特别是在缺乏专业人才的地区。然而，用于训练这些模型的敏感数据可能面临隐私攻击。攻击者利用成员推理攻击（MIA）来确定个人的数据是否被用于训练模型。通过此类攻击，可以推断出患者的医疗数据和私人信息。此前关于数据风险的研究主要基于整个数据集，并未考虑个体风险。

在这项研究中，德国慕尼黑工业大学的Moritz Knolle和同事开展了一项隐私审查，重点关注了个人隐私风险，结果发现医疗AI模型可能对个人数据贡献者构成隐私风险。

研究人员利用7个由真实临床数据（包括医学影像、心电图和电子健康记录）组成的大型数据集，确定了贡献数据的患者中最为脆弱的群体。他们发现，在个人层面，MIA针对的目标几乎毫无差错地被成功识别出来。在群体层面，在数据集中被识别为代表性不足的群体包括罕见病患者、少数族裔或社会经济地位较低的人群。随着被AI模型编码的独特数据增多，研究人员发现，这些群体和个人变得更加脆弱，且面临不成比例的隐私攻击风险。他们还发现，MIA攻击的成功率会随着模型容量和规模的增加而上升。

这些发现表明，诸如MIA之类的隐私攻击在个体层面的精准打击效果，比目前普遍认为的更为显著。研究人员总结称，隐私风险评估必须将个体风险纳入考量，并对易受攻击的模型提供进一步保护。（来源：中国科学报 赵熙熙）

相关论文信息：https://doi.org/10.1038/s41586-026-10688-0

作者：Moritz Knolle 来源：《自然》