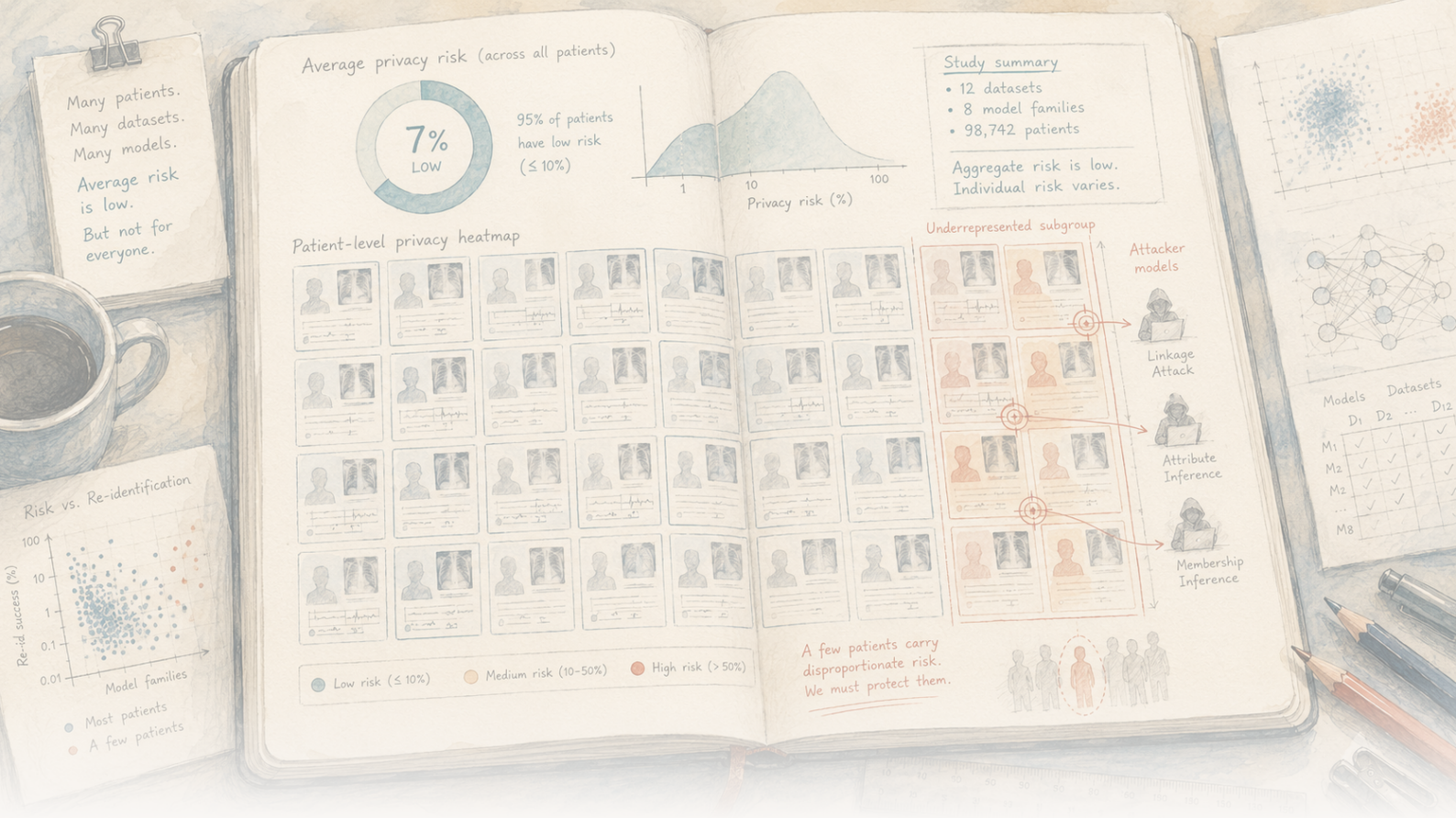


The
CLEAN PAPER

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

医疗 AI 可以在平均意义上是私密的，但仍然暴露特定的患者——尤其是那些最缺乏代表性的人

一个被称为「保护隐私」的医疗 *AI* 模型通常基于一个平均数字。这项研究认为，这一数字是错误的检验标准。作者研究成员身份推断攻击——即揭示特定人的记录是否在模型训练数据中，从而可能泄露其患有特定疾病——在七个医疗数据集（影像、心电图、电子记录）和众多模型上逐患者测量风险，而非汇总测量。模式是：在平均意义上看起来安全的模型，仍然可以让攻击者以近乎完美的准确率识别特定个体（攻击 $AUC \geq 0.95$）；被暴露的系统性地来自代表性不足的群体（少数族裔、罕见疾病、异常影像）；且随着模型的扩大而恶化。作者并非主张放弃医疗 *AI*——他们主张逐患者测量隐私、控制模型访问并使用差分隐私。令人不适的核心：「平均意义上是私密的」并非隐私保证，而它所辜负的患者已是那些最缺乏保护的人。

Written by Lucio Vaglio · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Disparate privacy risks from medical AI*, Moritz Knolle, Georg Kaissis, Daniel Rückert and colleagues (Technical University of Munich; Imperial College London; Hasso Plattner Institute), Nature (2026) · DOI: 10.1038/s41586-026-10688-0

隐私保证是对个人许下的承诺，而不是对平均值的承诺

当一个医疗人工智能模型被称为“保护隐私”时，这一说法通常建立在一个数字之上：在所有用于训练模型的患者中，推断某个人的数据是否属于训练集的平均概率很低。这听起来令人安心。但这篇论文提出了一个平静却令人不适的观点：这个数字选错了。

隐私不是平均值。它是对每个人作出的承诺——你的数据进入训练集，不会反过来暴露你。而平均值可能对几乎所有人兑现承诺，却对少数人彻底失信。研究人员以前所未有的分辨率，逐个患者测量隐私风险，发现的正是这一点：从总体上看似安全的模型，可能几乎完美地泄露特定个人是否属于训练集；而被泄露的个人，往往恰恰是本来就最缺乏保护的人。

作者做了什么

这支团队由莫里茨·克诺勒带领，成员包括慕尼黑工业大学的丹尼尔·吕克特、哈索·普拉特纳研究所的格奥尔格·凯西斯，以及伦敦帝国理工学院的同事。他们针对一种广为人知的隐私威胁——**成员推断攻击**——换了一种问法。与其问“这种攻击在整个数据集上的平均成功率是多少？”，他们问的是：“这个特定患者暴露的程度有多高？”

他们在**七个已有成熟应用的医学数据集**上进行了逐患者分析，数据类型差异很大，包括医学影像、心电图和电子健康记录；每个数据集又涵盖了 200 个模型。对于每个模型中的每位患者，他们估计攻击者有多大把握判断该患者的记录是否曾属于训练数据，随后按疾病状态、自报种族、保险、性别和影像采集协议等群体特征拆分结果。

What a membership inference attack actually is

这里的泄露比“模型把你的记录吐出来”更隐蔽。它涉及的是成员身份——你的数据曾经进入训练集这一单纯事实。

先看这类模型通常做什么。你给它一张扫描图，比如胸部 X 光片，它会返回概率：患肺炎的概率为 78%，同时给出心脏扩大、水肿和实变等情况的读数。这种日常诊断结果就是攻击所使用的全部信息，不需要任何特殊手段。

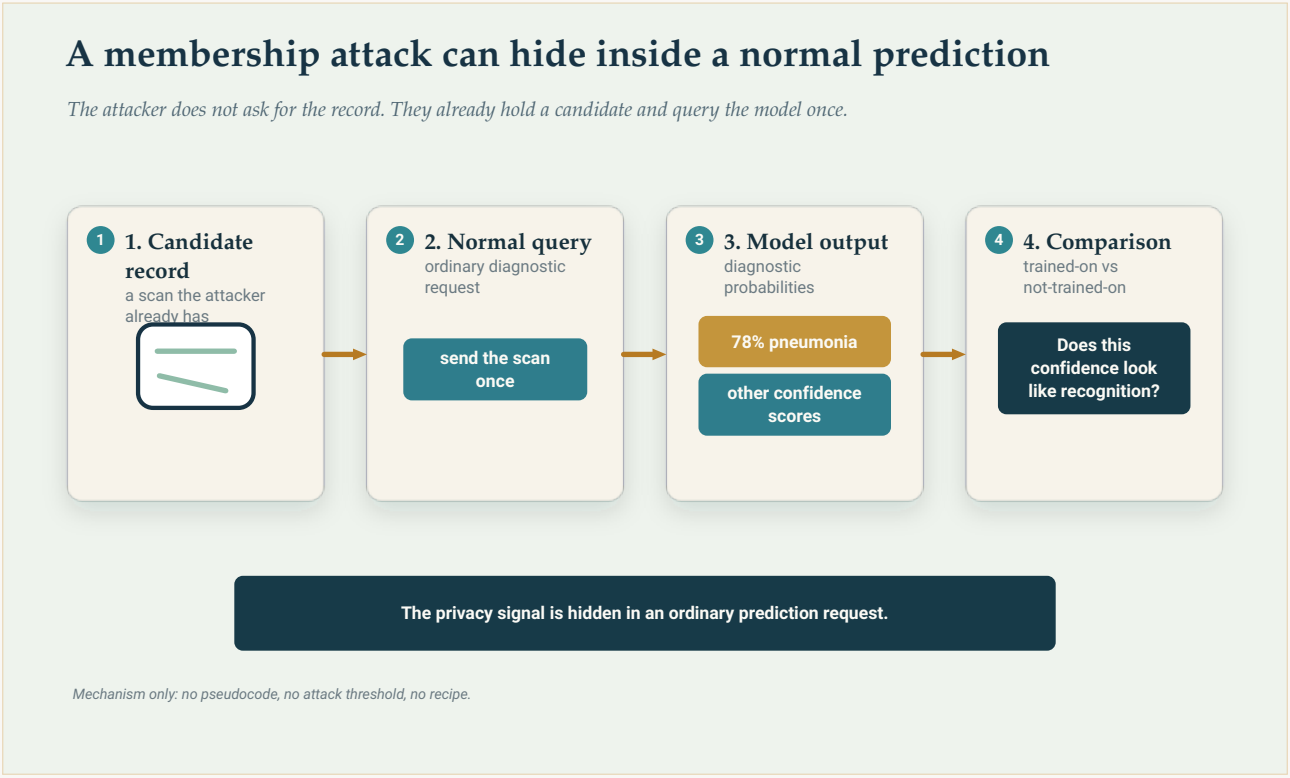
攻击利用的弱点是：对于自己训练过的确切样本，模型通常会比面对从未见过的样本**更有把握**。可以把它想成一个偷偷提前看过试卷的学生：遇到自己练习过的问题，答案会来得有点过快，也有点过于自信。根据这种破绽判断某条特定记录是否是模型学习过的样本，这就是“成员推断”的含义。

具体来说，攻击会这样进行。有人想知道某个人的扫描图是否被用来训练模型。他们**不会**向模型索要这条记录——他们手里已经有一张候选扫描图（可能是这个人的原图，也可能是接近的副本）。他们把它像普通诊断请求一样发送一次，并记录模型回答时有多大把握。接着，他们检查这种把握程度更像是训练过这张扫描图的模型，还是没训练过的模型。攻击者可以在普通电脑上自行训练一个替代模型（“参考模型”），以很低的成本学习这种特征性的过度自信，不需要特殊硬件。如果真实模型对这张扫描图异常确定——仿佛认出了它——这就是扫描图曾出现在训练数据中的有力证据。

令人不安的是，这个请求看起来完全不像攻击：它就是临床医生会发出的同一种诊断查询，而隐私信号藏在普通预测之中。这正是风险如此具体的原因——尽管到目前为止，这种风险是在明确说明的实验室假设下展示的，还没有在现实环境中被发现。

如果记录本身从未泄露，成员身份为什么重要？因为成员身份是一个事实。如果模型是在接受过某种癌症免疫疗法的患者数据上训练的，那么确认你的记录在其中，就会暴露出你很可能患过这种癌症——这是保险公司或雇主绝不应当推断出的信息。记录内容仍然被封存；逃出去的是“属于其中”这一事实。

作者明确列出了这些假设——能够访问模型的普通预测、拥有一条候选记录，以及攻击者自己的参考模型——不是为了提供操作指南，而是为了让人能够认真推演这一威胁，而不是把它一笔带过。



这种攻击不需要特殊的隐私 API。如果模型对自己以前见过的记录异常自信，那么一次普通的诊断查询就可能泄露成员身份信号。

他们发现了什么

共有三项发现，一项比一项尖锐。

平均值掩盖了暴露者。按总体情况测量——也就是通常的做法——这些模型中的许多看起来都具有令人安心的隐私性：对大多数患者而言，攻击并不比随机猜测更好。但逐患者观察呈现出另一番景象。正如

克诺勒在研究发布公告中所说，此前的评估“始终只测量所有患者的平均风险。我们首次在单个患者层面考察了风险——结果呈现出完全不同的图景。”对某些人来说，攻击的成功率**几乎达到完美**——作者将其测为攻击 AUC **0.95 或更高**（0.5 相当于抛硬币，1.0 则是完全准确）——即使整个数据集的平均结果看起来并不比随机猜测更好。

平均值可能接近随机水平，同时仍有一小部分记录暴露程度高得多。论文指出，隐私必须在患者层面进行检查，而不能只看总体情况。

Original Aurora diagram —The Clean Paper CC BY 4.0

暴露并不均等，而且落在代表性不足的人群身上。最容易受到近乎完美攻击的患者，系统性地来自**在数据中代表性不足的群体**：少数族裔、罕见疾病表型，以及影像特征不寻常的人。在电子记录数据集中，黑人患者出现在最脆弱记录中的频率比预期高出 **31%**；在乳腺摄影数据集中，被标记为疑似恶性肿瘤的扫描图在这一危险区域中的代表性高得惊人，达到 **+1,179%**。（这两个数字只是例子，并非全部情况——论文报告了多个群体中相同的偏斜——而且它们是小规模极端风险尾部中的相对过度代表：表示这些患者出现在暴露程度最高的记录中的频率高出多少，而不是黑人患者或疑似恶性乳腺摄影患者中有多少人暴露。）模型中可用于把这些患者“模糊”进去的相似样本更少，所以他们的记录更显眼——而显眼正是成员攻击所检测的东西。隐私失效并非随机发生；它集中在本来就处于边缘的人身上。

模型越大，情况越糟。随着模型容量增加，受到近乎完美攻击的患者数量急剧上升——在一个皮肤病学数据集中，这一比例从最小模型中基本为零，升至最大模型中的约**十分之一**。随着医疗人工智能模型变得更大、更强——这是整个领域正在前进的方向——作者警告，这种特定风险会变得更严重，而不是减轻。

吕克特的总结十分直白：“这种风险不可接受。健康数据高度敏感。”

为什么“平均来看安全”是错误的测试

人们很容易把较低的平均风险看成一份合格证明。论文的核心教训是，在这里取平均不只是有失精确——它测量的就是错误对象。

只有当隐私保证对风险最高的那个人也成立，而不只是对处于中间位置的人成立时，它才有意义。一个模型让 1,000 名患者中的 999 人无法被识别，却几乎完美地识别出其中一人；对这个人而言，无论如何都不能把它称作“99.9% 的隐私”。如果这个人恰好是罕见疾病患者或少数族裔患者，那么这个指标就不只是残缺不全，而是在不动声色地歧视。它报告的是多数人的安全，却把这称为所有人的安全。

这就是论文迫使我们完成的转变：从“这个模型平均有多隐私？”转向“哪个患者暴露得最多，他属于什么群体？”这是两个不同的问题，只有第二个才是真正的隐私问题。

这项研究没有证明——也没有声称——什么

- 它**没有**说应当放弃医疗人工智能。作者的框架是缓解风险，而不是退缩：测量并修复风险，不要停止构建。
- 它**不意味着**每个医疗模型都在泄露信息，也不意味着某个已部署的模型已经遭到攻击。它展示的是，风险确实存在，而且分布不均，标准的总体指标会漏掉这种风险。
- 它**不意味着**你的记录已经暴露。攻击需要特定条件——能够访问模型、拥有一条候选记录，以及攻击者自己的基础设施——并不是随手就能做到的事情。
- 它**没有**显示患者记录的内容会泄露。泄露的是成员身份——被纳入其中这一事实——危险来自另一种原因，而不是因为记录内容被直接泄露。
- 它**没有**把差异压缩成一个醒目的数字。强而且可复现的结果是这种模式——总体指标低估了个人风险，而代表性不足的患者承担了大部分风险——这一模式跨越了多个数据集和数百个模型。

证据有多强?

这是一个经验性、方法学上的结果，而且相当稳健。这一模式——总体隐私指标系统性低估逐患者风险，剩余风险集中在代表性不足的群体中，并且随着模型容量增加而恶化——在**七个数据类型不同的数据集**以及每个数据集中的大量模型上都成立。这样的广度，正是它成为可信测量结论、而不是单一数据集产物的原因。

这里有两个坦诚的限定。第一，这是对风险的展示，测量方式是运行作者自己构建的攻击；它刻画的是一个有能力的攻击者在既定假设下可能做到什么，而不是现实世界中的攻击发生得有多频繁。第二，那些醒目的数字——某些患者几乎完美的攻击成功率——刻意描述的是处境最糟的人；这正是研究的重点。因此，应当把它们理解为“尾部风险远比平均值所暗示的更重”，而不是“多数患者都暴露了”。

恰当的态度既不是惊慌，也不是置之不理：这是一项谨慎的多数据集研究，它表明我们用来认证医疗人工智能隐私的标准方式，对自身最糟糕的案例视而不见，而这些最糟糕的案例落在几乎没有缓冲余地的患者身上。

为什么这很重要

有两点让这不只是技术脚注。

第一，它改变了“保护隐私”应当被允许意味着什么。如果一个模型发布时带有平均隐私评分，这个评分完全可能确实很低，却仍然掩盖一部分能够被成员攻击近乎完美识别的患者。论文提出的实际要求很具体：在发布前**逐患者**评估隐私风险，控制谁能访问已部署的模型，并使用**差分隐私**等技术——在训练期间加入少量、经过仔细校准的噪声，削弱成员攻击，同时让模型的实用性付出真实且可管理的代价（隐私—效用权衡是明确存在的，并非没有成本）。“我们检查过平均值”不应再被算作完成了检查。

第二，它把隐私与公平交织在一起。在医学数据中代表性不足、因而已经受到医疗人工智能较差服务的同一批群体，结果发现也是这种人工智能保护其隐私最少的人。一个正在努力解决第一种不平等的领域，不能把第二种不平等当成别人的职责。面对的是同一批人。

医疗人工智能隐私那套令人安心的叙事——我们测过了，平均值很低，没问题——正是这篇论文要拆解的叙事。它不是为了吓退任何人，不是为了让人远离这项技术，而是要把标准推回应在的位置：只有检查过最可能受到伤害的那个人，才能声称自己兑现了隐私承诺。

简明总结

成员推断攻击试图确定某个特定人的记录是否在人工智能模型的训练数据中——而且因为成员身份本身可能暴露信息（比如你患过某种疾病），即使底层记录从未出现，这也是真实的隐私泄露。此前的研究测量的是这类攻击在整个数据集上的平均成功率。本研究则在七个医学数据集（影像、心电图、电子记录）和每个数据集的众多模型上，按患者进行测量，发现总体指标严重低估了风险：即使整个数据集的平均结果看起来安全，某些个人仍可能几乎完美地被识别（攻击 AUC ≥ 0.95）；最脆弱的系统性地来自代表性不足的群体（少数族裔、罕见疾病、不寻常的影像）；而且问题会随着模型规模扩大而加剧。作者并不主张放弃医疗人工智能——他们主张在个人层面测量隐私、控制模型访问，并使用差分隐私。要点是：“平均来看是私密的”不是隐私保证，而隐私保证未能保护的人通常正是本来就最缺乏保护的人。

不绕弯检查

论文显示了什么：在七个医学数据集和每个数据集的众多模型中，逐患者的成员推断风险远高于总体指标所显示的水平——攻击成功率最高可接近完美（AUC ≥ 0.95）——而且这种剩余风险不成比例地落在代表性不足的群体身上，并随着模型容量增加而上升。

什么是合理但尚未得到证明的：现实中的攻击者已经在利用这一点攻击已部署的临床模型；本研究在既定假设下展示了这种能力及其分布，但没有测量现实环境中攻击发生的频率。

研究没有显示什么：医疗人工智能应当被放弃；每个模型都在泄露信息，或某个特定的已部署模型已经被攻破；患者记录的内容（而不是成员身份）暴露；或者存在一个单一、量化的差异程度数字——稳健的结果是这种模式，而不是某一个数字。

主要局限：它通过作者自己的攻击测量最坏情况风险，而不是观察到的事件；那些戏剧性的数字刻意描述了暴露程度最高的个人；不同群体之间的具体差异是在多个数据集中作为一致模式呈现的，而没有被压缩成一个数字。

普通读者应当有多大信心？可以高度确信，总体隐私指标低估了个人风险，剩余风险集中在代表性不足的患者身上，而且这种风险会随着模型规模增大而恶化——这些结论在许多数据集和模型上都得到了展示。对于此类攻击在现实世界中的发生频率，信心应当是中等，因为本研究没有测量这一点。安全的理解方式不是“医疗人工智能会泄露你的数据”，也不是“隐私问题已经解决”，而是：“标准的隐私检查看不见自身最糟糕的案例，而这些案例落在最脆弱的患者身上——所以检查方式必须改变。”

来源

Knolle et al., Disparate privacy risks from medical AI, Nature (2026)
https://doi.org/10.1038/s41586-026-10688-0

Technical University of Munich —press release, ‘Study reveals privacy risks in medical AI’ (2026)
https://www.tum.de/en/news-and-events/all-news/press-releases/details/study-reveals-privacy-risks-in

Medical Xpress (2026) —coverage of the study
https://medicalxpress.com/news/2026-06-patient-groups-vulnerable-privacy-medical.html