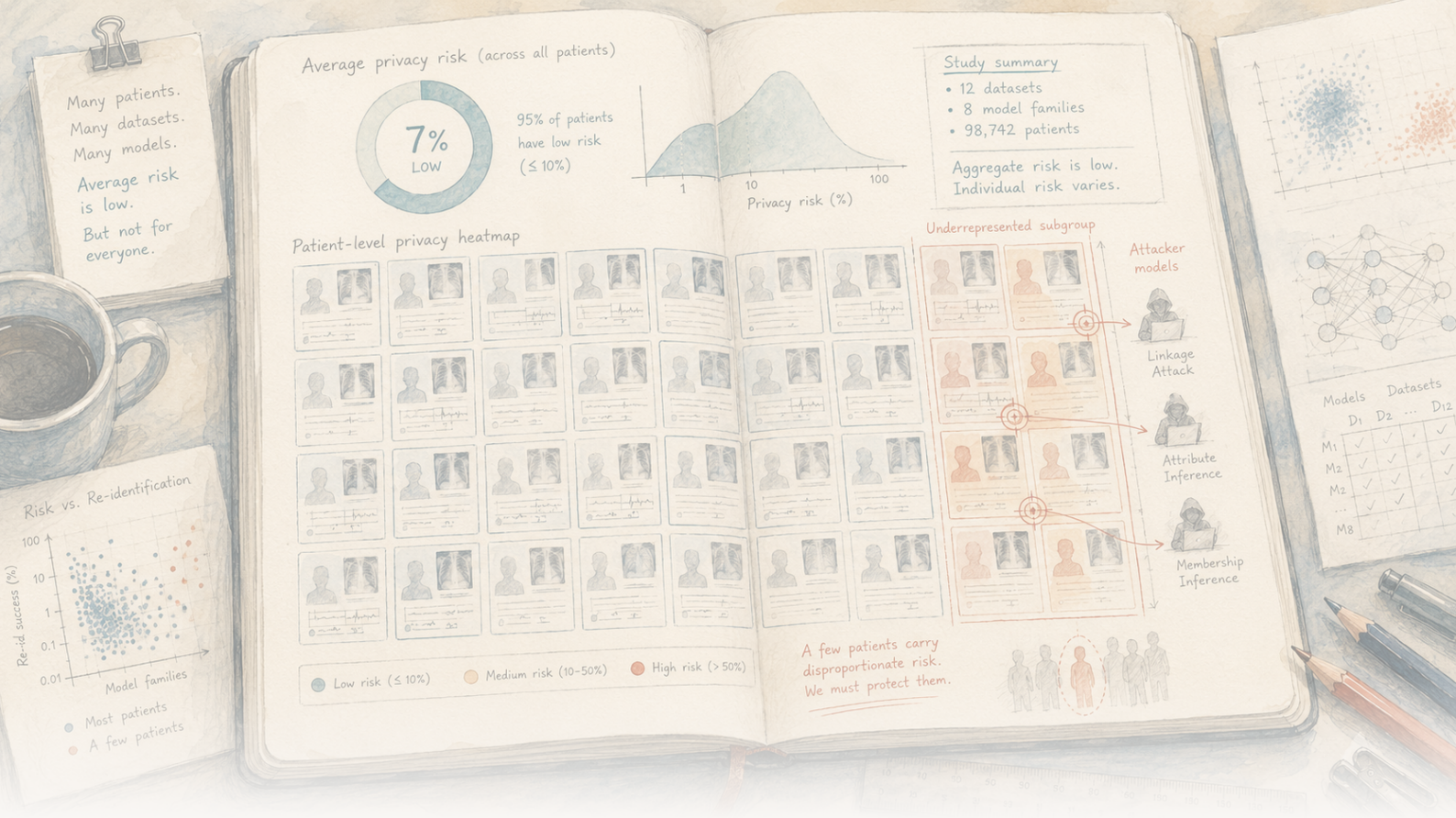


The CLEAN PAPER

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 · 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

隐私保证是对一个人的承诺，而不是对一个平均值的承诺

当一个医疗 AI 模型被称为” 隐私保护” 时，这一说法通常建立在一个数字之上：在所有为训练该模型贡献了数据的患者当中，任何个人的成员身份能被推断出来的平均概率很低。这听起来令人安心。而这篇论文那个安静而令人不安的观点是：这是一个错误的数字。

隐私不是一个平均值。它是对每一个个体作出的承诺——即身处训练集之中不会反过来暴露他们。而一个平均值可以对几乎所有人都信守这一承诺，却对少数几个人彻底违背它。研究者们以前所未有的分辨率、逐患者地衡量隐私风险，并恰恰发现了这一点：那些在汇总层面看起来很安全的模型，可能会近乎完美地泄露特定个体的成员身份——而且被它们泄露的那些个体，恰恰绝大多数是本就最缺乏保护的人。

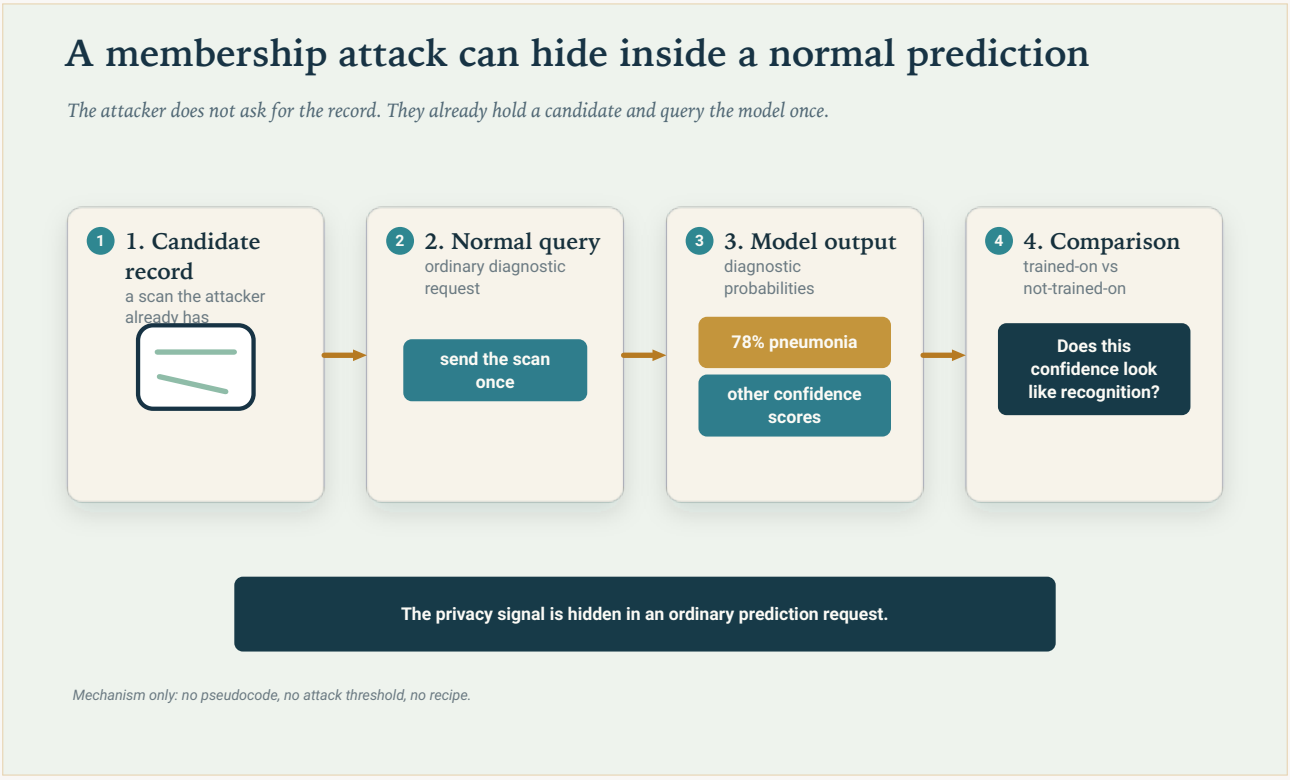
作者们做了什么

这个团队——由 Moritz Knolle 领衔，成员包括 Daniel Rückert（两人都在 Technical University of Munich）和 Georg Kaissis（Hasso Plattner Institute），以及 Imperial College London 的同事们——拿起了一种被称为**成员推断攻击**的著名隐私威胁，并改变了所提的问题。他们没有问” 平均而言，这种攻击在整个数据集上成功的频率有多高？“，而是问” 对于这一位特定的患者，他们暴露的程度有多深？ ”

他们在**七个成熟的医疗数据集**上运行了这项逐患者分析，这些数据集横跨了差异极大的数据类型——医学影像、心电图（ECG）以及电子健康记录——并且，在这些数据集上，每个数据集各训练了 200 个模型。对于每个模型中的每一位患者，他们估计了攻击者能以多大的把握判断出这个人的记录是否曾是训练数据的一部分，然后按群体拆解结果：疾病状态、自我报告的种族、保险、性别以及影像采集方案。

> **成员推断攻击究竟是什么**
>
> 这里的泄露比” 模型把你的记录吐出来” 要微妙得多。它关乎的是成员身份——仅仅是你的数据曾出现在训练集中这一事实。
> 先从这类模型平常做什么说起。你递给它一张扫描片——比如说一张胸部 X 光片——它回递给你一组概率：78% 的肺炎可能性，连同心脏肥大、水肿、实变的读数。那个日常的诊断答案就是攻击唯一会用到的东西。没有任何奇技淫巧。
> 它所依赖的弱点是这样的：一个模型对它所训练过的那些确切样本，通常会比对它从未见过的样本**略微更有把握**。想象一个偷偷提前看过考卷的学生——在他练习过的那些题目上，答案来得快了那么一点、笃定了那么一点。从这个破绽出发，推算出某一份特定的记录是否属于模型研习过的样本之一，这就是” 成员推断” 的含义。
> 那么这就是这场攻击，一步一步来。有人想知道某个特定的人的扫描片是否被用于训练模型。他们**并不**向模型索要那份记录——他们手里早已握有一张候选扫描片（那个人本人的，或一份接近的副本）。他们把它作为一次普通的诊断请求送进去一次，并记录下答案有多笃定。然后他们检查这份笃定看起来更像一个训练过这张扫描片的模型，还是一个没训练过的模型——这是一个他们可以廉价地做出的比较：只需在一台普通电脑上、无需任何特殊硬件，训练一个自己

的替身（一个”参考模型”），来学习那种泄露天机的过度自信长什么样。如果真实的模型对这张扫描片异乎寻常地确信——仿佛认得它——那就是这张扫描片曾在训练数据中的有力证据。
令人不安之处在于，这次请求没有任何地方看起来像一场攻击：它和临床医生会发出的诊断查询一模一样，而隐私信号就藏在一个普通预测的内部。这正是为何这种风险是具体实在的——尽管到目前为止，它是在明确陈述的实验室假设下被演示出来的，而非在现实世界中被逮个正着。
如果记录本身从不泄露，成员身份为什么重要？因为成员身份本身就是一个事实。如果一个模型是在接受过某种特定癌症免疫疗法的患者身上训练出来的，那么确认你的记录在其中，就揭示了你很可能患过那种癌症——这正是保险公司或雇主永远不该能够推断出来的那类事情。内容始终被封存；逃逸出来的是”属于其中”这一事实。
作者们明明白白地列出了这些假设——能访问模型的普通预测、一份候选记录，以及攻击者自己的参考模型——这并非作为一份操作指南，而是为了让这一威胁能被推理审视，而非被含糊搪塞。

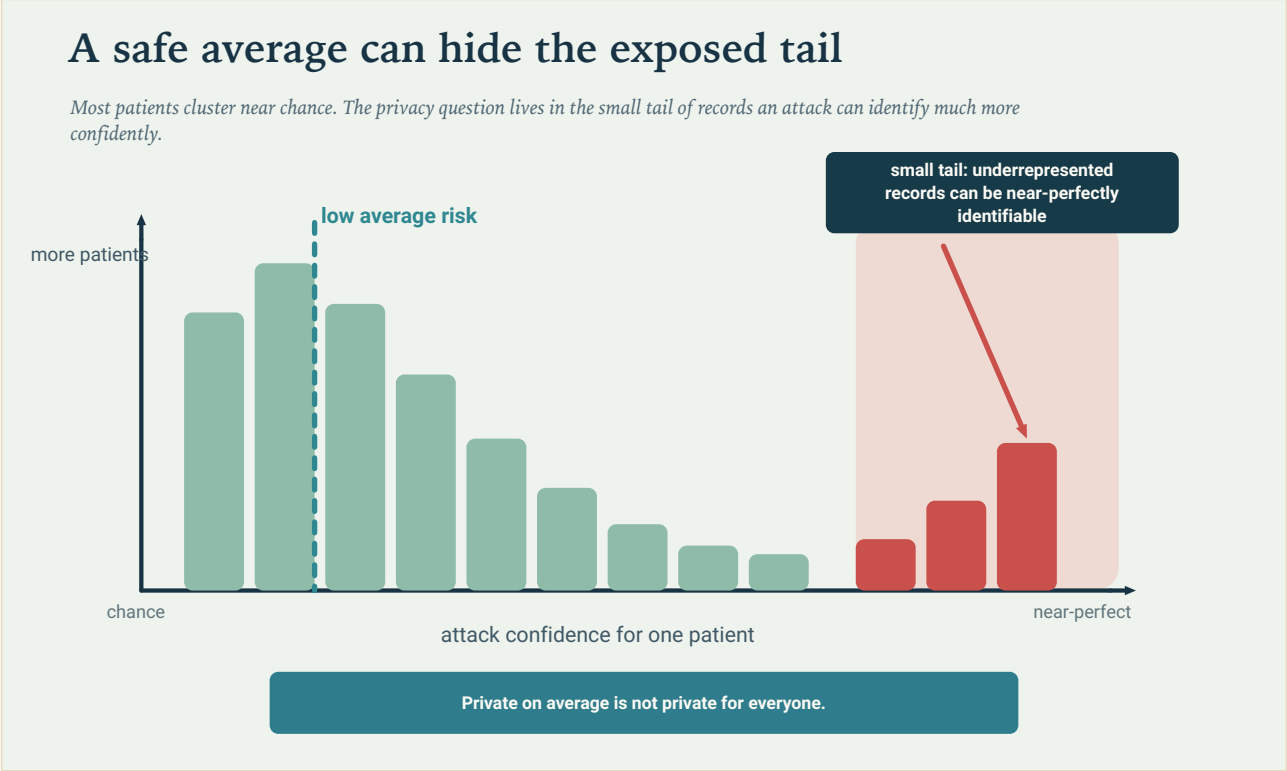


这场攻击不需要一个特殊的隐私 API。如果模型对一份它此前见过的记录异乎寻常地自信，一次正常的诊断查询就可能泄露出一个成员身份信号。

他们发现了什么

三项发现，一项比一项更为尖锐。

平均值掩盖了那些被暴露的人。以汇总方式衡量——也就是惯常的做法——这些模型中的许多看起来私密得令人安心：对大多数患者而言，攻击的表现并不比随机猜测更好。而逐患者的视角讲述了一个不同的故事。正如 Knolle 在这项研究的公告中所说的那样，以往的评估" 始终只衡量了所有患者的平均风险。我们首次在个体患者的层面上审视了风险——它描绘出一幅截然不同的图景。"对于某些个体，攻击成功率**近乎完美**——作者们将其衡量为一个 **0.95 或更高**的攻击 AUC（0.5 是掷硬币的水平，1.0 是毫无瑕疵）——而与此同时，全数据集范围内的平均值看起来并不比随机猜测更好。



平均值可以停留在接近随机猜测的水平，而一小撮尾部的记录仍然要暴露得多。这篇论文的观点是：隐私必须在患者层面而不仅仅是在汇总层面接受检验。

Original Aurora diagram —The Clean Paper CC BY 4.0

这种暴露是不均等的——而且它落在了代表性不足的群体身上。最容易遭受近乎完美攻击的患者，系统性地都是那些处于在**数据中代表性不足**的群体中的人：少数族裔、罕见病表型、异常的影像特征。在电子记录数据集中，Black patients 在最脆弱的那批记录里出现的频率比预期高出 **31%**；在乳腺 X 线摄影数据集中，被标记为疑似恶性的扫描片在那个危险地带里过度代表的幅度惊人，达到了 **+1,179%**。（这两个数字是例子，而非全貌——论文报告了在若干群体中同样的偏斜——而且它们是那个极小的极端风险尾部之内的相对过度代表：指的是这些患者在最被暴露的记录中出现的频率高出多少，而不是 Black patients 或疑似乳腺摄影患者中被暴露的那一比例。）一个模型能用来把这些患者模糊融入其中的相似样本更少，所以他们的记录会凸显出来——而凸显出来恰恰是成员推断攻击所要探测的东西。这种隐私失效并非随机；它集中落在那些本就处于边缘的人身上。

更大的模型让情况更糟。暴露于近乎完美攻击之下的患者数量，随着模型容量的增大而急剧上升——在一个皮肤病学数据集中，这一比例从最小模型中的基本为零，攀升到了最大模型中的大约**十分之一**。随着医疗 AI 模型变得越来越大、越来越强——这正是整个领域正在前进的方向——作者们警告说，这个特定的风险会变得更严重，而非更轻。

Rückert 的总结毫不含糊：“这不是一个可以容忍的风险。健康数据高度敏感。”

为什么” 平均而言安全” 是错误的检验标准

人们很容易把一个较低的平均风险读作一张健康证明。这篇论文的核心教训是：这里的求平均不仅仅是不够精确——它衡量的根本就是错误的东西。

一项隐私保证只有在它对风险最大的那个人也成立时才有意义，而不是对处于中间位置的那个人成立就够了。一个 1000 名患者中有 999 名无法被还原、但有 1 名能被近乎完美地识别出来的模型，在任何对那 1 个人有意义的层面上都不是”99.9% 私密”——而如果那 1 个人可以预料地就是罕见病患者或少数族裔患者，那么这个指标就不只是不完整，它是在悄悄地施行歧视。它报告的是多数人的安全，却把它称作所有人的安全。

这就是这篇论文所迫使的转变：从这个模型平均而言有多私密？转向谁是最被暴露的患者，而他们又是谁？这是两个不同的问题，而只有第二个才是一个隐私问题。

这项研究没有证明——或声称——什么

- 它**没有**说医疗 AI 应该被抛弃。作者们的立场是缓解，而非撤退：衡量并修复这个风险，而不是停止建设。
- 它**并不**意味着每一个医疗模型都在泄露，或者任何一个已部署的模型已经遭到攻击。它表明的是风险确实存在、且分布不均，以及标准的汇总指标会漏掉它。
- 它**并不**意味着你的记录已经被暴露了。这场攻击需要特定的条件——能访问模型、一份候选记录，以及攻击者自己的基础设施——而不是一种随手可得的能力。
- 它**并没有**表明患者的内容会泄露。泄露出来的是成员身份——被纳入其中这一事实——它之所以危险是出于一个不同的原因，而不是因为记录被倾倒了出来。
- 它**并没有**把这些差距归结为一个单一的醒目数字。那个强有力、可复现的结果是这个模式——汇总指标低估了个体风险，而代表性不足的患者承受了其中的大部分——它横跨多个数据集和数百个模型。

证据有多强?

这是一个实证的、方法学上的结果，而且是一个稳健的结果。这个模式——汇总隐私指标系统性地低估了逐患者风险，残余风险集中在代表性不足的群体身上，且它随模型容量而恶化——在**七个不同数据类型的数据集**和每个数据集大量的模型上都成立。正是这种广度，让一个测量层面的论断变得可信，而不是沦为单一数据集的假象。

两点诚实的告诫。第一，这是对风险的一次演示，是通过运行作者们自己构建的攻击来衡量的；它刻画的是一个有能力的攻击者在既定假设下能够做到多好，而不是现实世界中的攻击实际发生得有多频

繁。第二，那些鲜明的数字——某些患者近乎完美的攻击成功率——按其设计描述的就是境况最糟的那批个体；这正是全部的用意所在，它们应被读作" 尾部远比平均值所暗示的沉重得多"，而不是" 大多数患者都被暴露了"。

恰当的立场既不是惊慌也不是不屑：一项审慎的、多数据集的研究表明，我们认证医疗 AI 隐私的标准做法对它自己最糟糕的那些情形是盲目的——而那些最糟糕的情形，恰恰落在了最没有余地可供腾挪的患者身上。

为什么这很重要

有两件事让这不止是一个技术脚注。

第一，它改变了" 隐私保护" 这个说法应该被允许指代什么。如果一个模型带着一个平均隐私分数被发布出来，那个分数可以是货真价实地低，却依然掩藏了一部分能被成员推断攻击近乎完美地识别出来的患者。这篇论文的实际诉求是具体的：在发布之前**逐患者地**评估隐私风险，控制谁能访问已部署的模型，并使用诸如**差分隐私**这样的技术——在训练过程中加入的、经过小心校准的少量噪声，它能钝化成员推断攻击，代价是模型的有用性会有一份真实且受控的损失（隐私与效用之间的权衡是明摆着的，而非免费的）。“我们检查了平均值" 应该不再算作已经检查过了。

第二，它把隐私和公平交织在了一起。那些在医疗数据中代表性不足的群体——因此本就被医疗 AI 服务得更差——结果恰恰就是他们的隐私被那个 AI 保护得最少的群体。一个正在为第一重不平等而努力的领域，不能把第二重当作别人部门的事。他们是同一批人。

那个关于医疗 AI 隐私的令人安心的故事——我们衡量过了，平均值很低，我们没问题——正是这篇论文要拆解的故事。这样做不是为了把任何人吓离这项技术，而是为了把标准挪到它本该在的地方：一个只有当你已经为最可能受到伤害的那个人检验过它之后，你才能声称自己信守的承诺。

干净的总结

成员推断攻击试图判定某个特定的人的记录是否曾在一个 AI 模型的训练数据之中——而由于成员身份本身可能就带有揭示性（比如说，揭示你患过某种特定的疾病），所以即便底层的记录从未浮出水面，这也是一次货真价实的隐私泄露。此前的工作衡量的是这类攻击平均而言在一个数据集上成功的频率。这项研究衡量的是它逐患者的情况，横跨七个医疗数据集（影像、ECG、电子记录）以及每个数据集的许多模型，并发现汇总指标严重低估了风险：某些个体能被近乎完美地识别出来（攻击 AUC ≥ 0.95），即便全数据集范围的平均值看起来很安全；最脆弱的那些人系统性地来自代表性不足的群体（少数族裔、罕见病、异常影像）；而且这个问题随模型规模的增大而增长。作者们并不主张抛弃医疗 AI——他们主张在个体层面衡量隐私、控制模型访问权限，并使用差分隐私。要点在于：“平均而言私密" 并不是一项隐私保证，而它辜负的那些人，通常就是本就最缺乏保护的人。

无废话核查

这篇论文表明了什么：横跨七个医疗数据集以及每个数据集的许多模型，逐患者的成员推断风险对某些个体而言远高于汇总指标所暗示的水平——高达近乎完美的攻击成功率（AUC $\geq$ 0.95）——而这份残余风险不成比例地落在代表性不足的群体身上，并随模型容量的增大而增长。

什么是貌似合理但未经证实的：现实世界中的攻击者已经在针对已部署的临床模型利用这一点；这项研究演示的是既定假设下的能力及其分布，而不是现实世界中攻击的发生频率。

它没有表明什么：医疗 AI 应该被抛弃；每个模型都在泄露，或任何一个特定的已部署模型已被攻破；患者记录的内容（而非成员身份）被暴露；一个单一的量化差距数字——那个稳健的结果是这个模式，而不是某一个数字。

主要的局限：它是通过作者们自己的攻击来衡量最坏情形风险的，而不是通过观测到的事件；那些戏剧性的数字按其设计描述的是最被暴露的那批个体；而各群体之间确切的差距，是作为一个横跨多个数据集的一致模式呈现出来的，而非被归结为一个数字。

一位普通读者应该抱有多大的信心？对以下几点信心很高：汇总隐私指标低估了个体风险，残余风险集中在代表性不足的患者身上，以及这一点随模型规模的增大而恶化——它是在许多数据集和模型上被演示出来的。对这类攻击在现实世界中的发生频率则只有中等信心，因为这项研究并没有衡量它。稳妥的读法是：既不是” 医疗 AI 会泄露你的数据”，也不是” 隐私问题已经解决”，而是” 标准的隐私核查对它自己最糟糕的那些情形是盲目的，而那些情形落在了最脆弱的患者身上——所以这套核查必须改变。”

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-ai

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