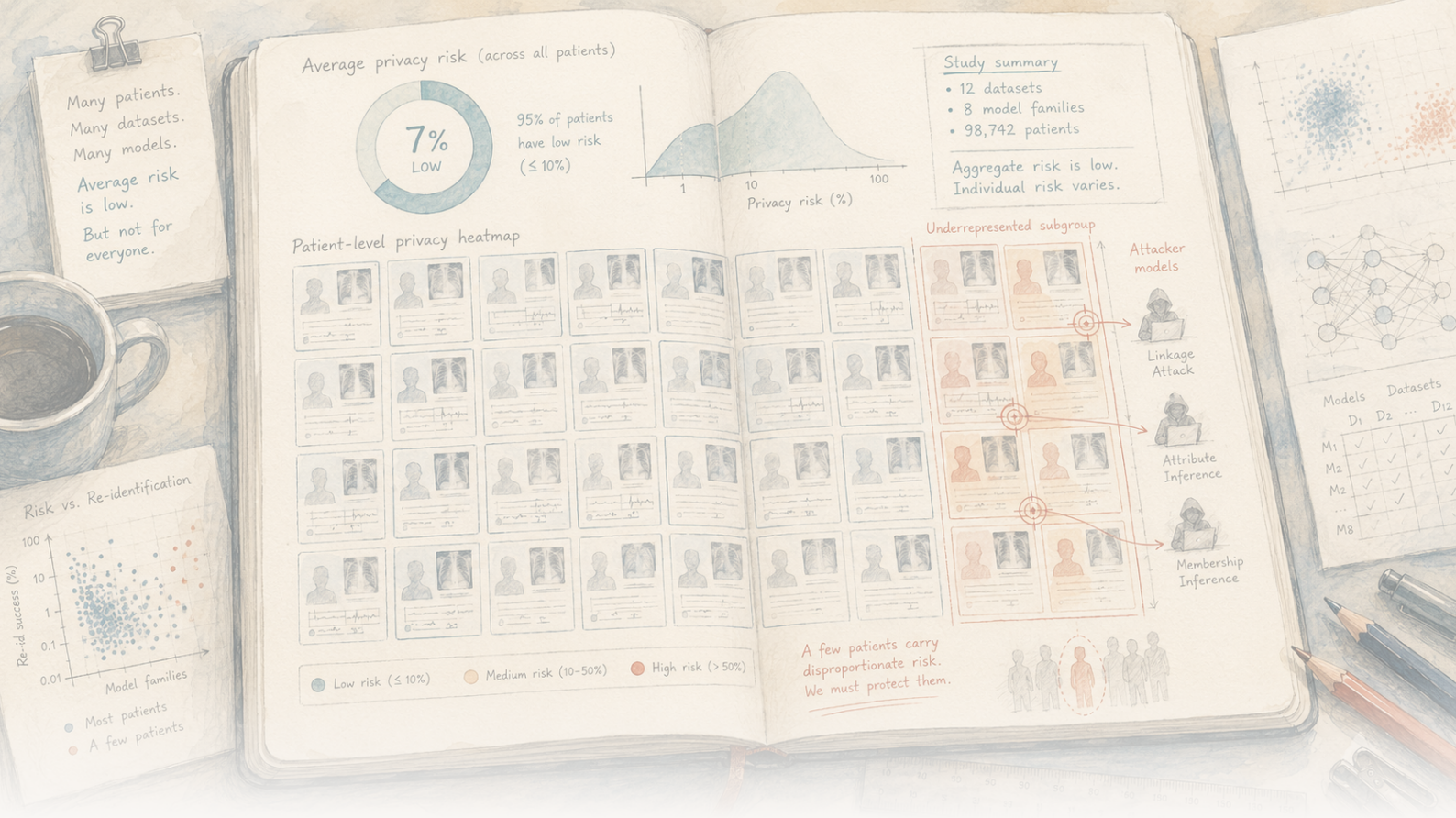


The CLEAN PAPER

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

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

一個被稱為「保護隱私」的醫療 *AI* 模型通常基於一個平均數字。這項研究認為，這一數字是錯誤的檢驗標準。作者研究成員身份推斷攻擊——即揭示特定人的記錄是否在模型訓練數據中，從而可能洩露其患有特定疾病——在七個醫療數據集（影像、心電圖、電子記錄）和眾多模型上逐患者測量風險，而非匯總測量。模式是：在平均意義上看起來安全的模型，仍然可以讓攻擊者以近乎完美的準確率識別特定個體（攻擊 $AUC \geq 0.95$）；被暴露的系統性地來自代表性不足的群體（少數族裔、罕見疾病、異常影像）；且隨著模型的擴大而惡化。作者並非主張放棄醫療 *AI*——他們主張逐患者測量隱私、控制模型訪問並使用差分隱私。令人不適的核心：「平均意義上是私密的」並非隱私保證，而它所辜負的患者已是那些最缺乏保護的人。

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

隱私保證是對個人的承諾，而不是對平均值的承諾

當醫療 AI 模型被稱為「保護隱私」時，這項說法通常建立在一個數字上：在所有資料曾用於訓練模型的患者中，平均而言，任何一人的成員身分被推斷出的機率都很低。這聽起來令人安心。但這篇論文提出了一個低調卻令人不安的觀點：這是錯的數字。

隱私不是平均值。它是對每一個人作出的承諾——一個人出現在訓練集裡，不會反過來使其身分曝光。而平均值可以對幾乎所有人守住這個承諾，卻對少數人徹底失守。研究人員逐一測量每位患者的隱私風險，達到前所未有的解析度，並確實發現了這一點：整體看似安全的模型，可能幾乎完美地洩漏特定個人的成員身分——而且身分遭洩漏的個人，往往正是原本最缺乏保障的人。

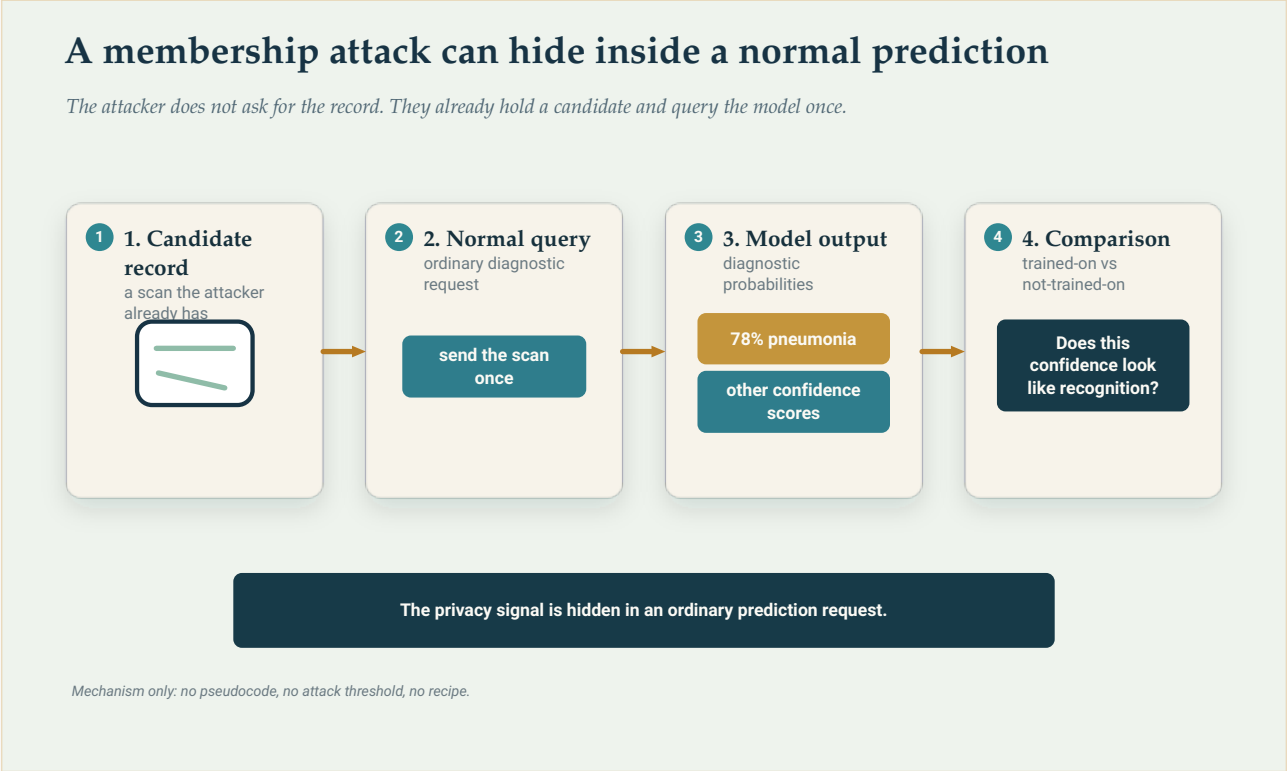
作者做了什麼

這個團隊由 Moritz Knolle 領導，成員包括 Daniel Rückert（兩人皆任職於慕尼黑工業大學）與 Georg Kaissis（任職於 Hasso Plattner Institute），以及倫敦帝國學院的同事。他們採用一種廣為人知、稱為**成員推斷攻擊**的隱私威脅，並改變了問題的問法。他們不再問：「這種攻擊在整個資料集上的平均成功率是多少？」而是問：「對這位特定患者而言，暴露程度有多高？」

他們在**七個成熟的醫療資料集**上進行逐一患者的分析，涵蓋截然不同的資料類型——醫學影像、心電圖與電子健康紀錄——而且每個資料集都分析了 200 個模型。對每個模型中的每位患者，他們估計攻擊者能有多大把握判斷該患者的紀錄是否曾納入訓練資料，接著按群組拆分結果：疾病狀態、自我報告的種族、保險、性別與影像協定。

> **成員推斷攻擊到底是什麼**
>
> 這裡的洩漏比「模型吐出你的紀錄」更微妙。它關乎成員身分——僅僅是你的資料曾在訓練集裡這個事實。
>
> 先從這類模型通常做的事開始。你交給它一張掃描影像——例如胸部 X 光——它會回傳機率：有 78% 的機率是肺炎，並同時給出心臟肥大、水腫與實變的判讀。這個日常診斷回覆就是攻擊唯一使用的東西，不需要任何特殊手段。
>
> 攻擊所利用的弱點是：模型通常對自己訓練過的確切範例，比對從未見過的範例更**有把握**。可以想像一名偷偷提前看過考卷的學生——遇到自己練習過的題目時，答案會快得有點不自然，也會自信得有點過頭。根據這個破綻，判斷某筆紀錄是否是模型學習過的範例之一，這就是「成員推斷」的意思。
>
> 以下是這種攻擊逐步運作的方式。有人想知道某個人的掃描影像是否被用來訓練模型。他們**不會**向模型索取那筆紀錄——他們手上已經有一張候選掃描影像（可能是那個人自己的，也可能是極為相近的副本）。他們把影像送入模型一次，像平常一樣提出診斷請求，並記下模型有多大把握。接著，他們檢查這個信心程度看起來更像哪一種模型：曾經用這張掃描影像訓練過的模型，還是沒有用過的模型。這項比較可以藉由自行訓練一個替代模型（「參考模型」）來低成本完成，讓它在一般電腦上、無需特殊硬體，學會辨認那種洩漏線索的過度自信。如果真實模型對這張掃描影像異常確定——彷彿認得它——那就是這張影像曾出現在訓練資料中的有力證據。

> 令人不安之處在於，這個請求看起來完全不像攻擊：它和臨床醫師會提出的診斷查詢相同，而隱私訊號藏在一般的預測裡。這正是風險如此具體的原因——儘管截至目前，這種風險是在明確陳述的實驗室假設下得到展示，而不是在真實環境中被發現。
> 如果紀錄本身從未洩漏，成員身分為什麼重要？因為成員身分是一項事實。如果模型是用接受過某種癌症免疫療法的患者訓練的，那麼確認你的紀錄在其中，就會揭示你很可能罹患過那種癌症——這是保險公司或雇主絕不應該能推斷出的事情。內容仍然封存；洩漏出去的是「屬於其中」這個事實。
> 作者清楚列出這些假設——能存取模型的一般預測、一筆候選紀錄，以及攻擊者自己的參考模型——不是要提供操作指南，而是讓人能夠推究這項威脅，而不是含糊帶過。



這種攻擊不需要特殊的隱私 API。如果模型對曾經看過的紀錄異常有把握，一次普通的診斷查詢就可能洩漏成員身分訊號。

他們發現了什麼

三項發現，每一項都比前一項更加尖銳。

平均值掩蓋了暴露者。按整體測量——也就是通常的做法——許多模型看起來很能保護隱私，令人安心：對大多數患者而言，攻擊的效果不比隨機猜測好。逐一患者的視角卻呈現出不同的景象。正如 Knolle 在研究公告中所說，先前的評估「一直只測量所有患者的平均風險。我們首次在個別患者的層

級檢視風險——結果呈現出截然不同的圖像。」對某些個人而言，攻擊的成功率**幾乎完美**——作者將其測量為攻擊 AUC **0.95 或更高**（0.5 代表擲硬幣般的隨機猜測，1.0 代表完全無誤）——即使整個資料集的平均表現看起來也不比隨機猜測好。

平均值可以接近隨機猜測的程度，但仍有一小部分紀錄的暴露程度高得多。這篇論文的重點是，隱私必須在患者層級檢查，而不能只看整體。

Original Aurora diagram —The Clean Paper CC BY 4.0

暴露並不平等——而且落在代表性不足的群體身上。最容易受到近乎完美攻擊的患者，系統性地來自**在資料中代表性不足的群體**：少數族裔、罕見疾病表型，以及不常見的影像特徵。在電子紀錄資料集中，黑人患者出現在最脆弱紀錄中的頻率，比預期高出 **31%**；在乳房攝影資料集中，被標記為疑似惡性腫瘤的掃描影像，在這個危險區域的過度代表程度更達到驚人的 **+1,179%**。（這兩個數字只是例子，而非完整圖景——論文也報告了數個群體中的相同偏斜——而且它們是小型極端風險尾部中的相對過度代表：指這些患者出現在暴露程度最高的紀錄中的頻率高出多少，而不是黑人患者或乳房攝影疑似惡性患者中有多少比例遭到暴露。）模型中可用來讓這些患者融入其中的相似範例較少，因此他們的紀錄會凸顯出來——而凸顯正是成員推斷攻擊所偵測的特徵。這項隱私失效並非隨機；它集中在本已處於邊緣的人身上。

模型越大，情況越糟。隨著模型容量增加，容易受到近乎完美攻擊的患者數量急劇上升——在一個皮膚科資料集中，這個比例從最小模型中幾乎為零，攀升到最大模型約**十分之一**。作者警告，隨著醫療 AI 模型變得更大、更有能力——這正是整個領域前進的方向——這項特定風險會變得更嚴重，而不是減輕。

Rückert 的總結很直接：「這不是可以容忍的風險。健康資料高度敏感。」

為什麼「平均而言安全」是錯誤的測試

人們很容易把低平均風險解讀成一份健康證明。這篇論文的核心教訓是，在這裡取平均不只是精確度不足——它測量的是錯誤的東西。

隱私保證只有在對風險最高的人也成立時才有意義，而不是只對處於中間位置的人成立。如果 1,000 名患者中有 999 人無法被識別，卻有一人幾乎可以被完美識別，那麼對那個人而言，這在任何有意義的層面上都不能稱為「99.9% 的隱私」——而且如果那一人往往就是罕見疾病患者或少數族裔患者，那個指標就不只是片面，而是悄然帶有歧視性。它報告多數人的安全，卻把那稱作所有人的安全。

這就是論文迫使我們做出的轉變：從這個模型平均而言有多能保護隱私？轉向暴露程度最高的患者是誰，他們又是什麼樣的人？這是不同的問題，而只有第二個才是隱私問題。

這項研究沒有證明——或聲稱——什麼

- 它**沒有**說醫療 AI 應該被放棄。作者的立場是降低風險，而不是退縮：測量並修正風險，不要停止開發。
- 它**不代表**每個醫療模型都在洩漏資料，也不代表任何特定的已部署模型曾遭到攻擊。它顯示的是，風險確實存在，而且分布不均，以及標準的整體指標會漏掉這項風險。
- 它**不代表**你的紀錄已經暴露。這種攻擊需要特定條件——能存取模型、一筆候選紀錄，以及攻擊者自己的基礎設施——並非隨手就能做到。
- 它**沒有**顯示患者紀錄的內容會洩漏。洩漏的是成員身分——被納入其中這個事實——其危險另有原因，並不是因為紀錄被整筆倒出來。
- 它**沒有**把差異縮減成一個醒目的數字。強而且可重現的結果是這個模式——整體指標低估個人風險，而代表性不足的患者承擔了其中最大的部分——這個模式在多個資料集與數百個模型中都存在。

證據有多強？

這是一項實證性、方法論上的成果，而且相當穩健。這個模式——整體隱私指標系統性低估逐一患者的風險、剩餘風險集中在代表性不足的群體，以及風險隨模型容量增加而惡化——在**七個不同資料類型的資料集**與每個資料集的大量模型中都成立。正是這樣的廣度，使這項測量主張可信，而不是某個單一資料集造成的假象。

有兩點需要如實說明。第一，這是對風險的示範，測量方式是執行作者自己建立的攻擊；它描述的是有能力的攻擊者在明確陳述的假設下能做到多好，而不是真實世界的攻擊發生得多頻繁。第二，那些鮮明的數字——某些患者面臨近乎完美的攻擊成功率——是刻意描述處境最差的個人；這正是整項研究的重點，應該解讀為「尾部比平均值所暗示的沉重得多」，而不是「大多數患者都暴露了」。

適當的態度既不是恐慌，也不是不屑一顧：這是一項謹慎、涵蓋多個資料集的研究，顯示我們用來認證醫療 AI 隱私的標準方式，看不見自身最糟糕的案例——而且這些最糟糕的案例落在最沒有餘裕的患

者身上。

為什麼重要

有兩件事讓這不只是技術上的附帶說明。

第一，它改變了「保護隱私」應該被允許代表什麼。如果一個模型連同平均隱私分數一起發布，那個分數可能確實很低，卻仍然掩蓋一群能被成員推斷攻擊近乎完美識別的患者。這篇論文提出的實務要求很具體：在發布前**逐一患者**評估隱私風險，控制誰能存取已部署的模型，並使用**差分隱私**等技術——在訓練期間加入少量、經仔細校準的雜訊，以削弱成員推斷攻擊，但模型的實用性會付出真實且需要管理的代價（隱私—效用取捨是明確存在的，並非毫無代價）。「我們檢查過平均值」不應再被算作已經完成檢查。

第二，它把隱私與公平交織在一起。在醫療資料中代表性不足、因而本來就較難獲得醫療 AI 良好服務的群體，結果也是這種 AI 對其隱私保護最少的群體。一個努力處理第一種不平等的領域，不能把第二種不平等當成別人的責任。這是同一群人。

醫療 AI 隱私那個令人安心的故事——我們測量過了，平均值很低，所以沒問題——正是這篇論文所拆解的故事。目的不是嚇得任何人遠離這項技術，而是把標準移到它應在的位置：只有檢查過最可能受傷害的人，才能聲稱自己守住了這項承諾。

重點摘要

成員推斷攻擊試圖判定某個特定個人的紀錄是否出現在 AI 模型的訓練資料中——而且由於成員身分本身可能揭示資訊（例如你曾患有某種疾病），即使底層紀錄從未現身，這仍是真實的隱私洩漏。過去的研究測量這類攻擊在資料集上的平均成功率。本研究在患者層級測量這項風險，涵蓋七個醫療資料集（影像、ECG、電子紀錄）及每個資料集中的許多模型，並發現整體指標嚴重低估風險：即使整個資料集的平均結果看似安全，某些個人仍能被幾乎完美地識別（攻擊 AUC ≥ 0.95）；最脆弱者系統性地來自代表性不足的群體（少數族裔、罕見疾病、不常見的影像特徵）；而且問題會隨模型規模增大。作者並不主張放棄醫療 AI——他們主張在個人層級測量隱私、控制模型存取，並使用差分隱私。重點是：「平均而言能保護隱私」不是隱私保證，而它辜負的人，通常正是原本最缺乏保障的人。

務實檢視

論文顯示了什麼：在七個醫療資料集及每個資料集的許多模型中，逐一患者測得的成員推斷風險，對某些個人而言遠高於整體指標所暗示的程度——攻擊成功率最高可接近完美（AUC ≥ 0.95）——而且這項剩餘風險不成比例地落在代表性不足的群體身上，並隨模型容量增加。

合理但尚未證明的是什麼：真實世界的攻擊者已經在利用這項能力攻擊已部署的臨床模型；本研究是在明確陳述的假設下展示這項能力及其分布，而不是測量真實環境中的攻擊頻率。

它沒有顯示什麼：醫療 AI 應該被放棄；每個模型都在洩漏資料，或任何特定的已部署模型已遭突破；患者紀錄的內容（而非成員身分）暴露；或一個單一的量化差異數字——穩健的結果是這個模式，而不是一個數字。

主要限制：它透過作者自己的攻擊測量最壞情況下的風險，而不是觀察到的事件；引人注目的數字刻意描述暴露程度最高的個人；確切的群組差異在不同資料集中呈現為一致的模式，而不是被縮減成一個數字。

一般讀者應該有多大的信心？對於整體隱私指標低估個人風險、剩餘風險集中在代表性不足的患者身上，以及這種風險隨模型規模增大而惡化，可以有高度信心——這些現象已在許多資料集與模型中得到證明。至於這類攻擊在真實世界中的發生頻率，則應抱持中等信心，本研究並未測量這一點。穩妥的解讀不是「醫療 AI 會洩漏你的資料」，也不是「隱私問題已經解決」，而是「標準的隱私檢查看不見自身最糟糕的案例，而這些案例落在最脆弱的患者身上——因此檢查方式必須改變。」

來源

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