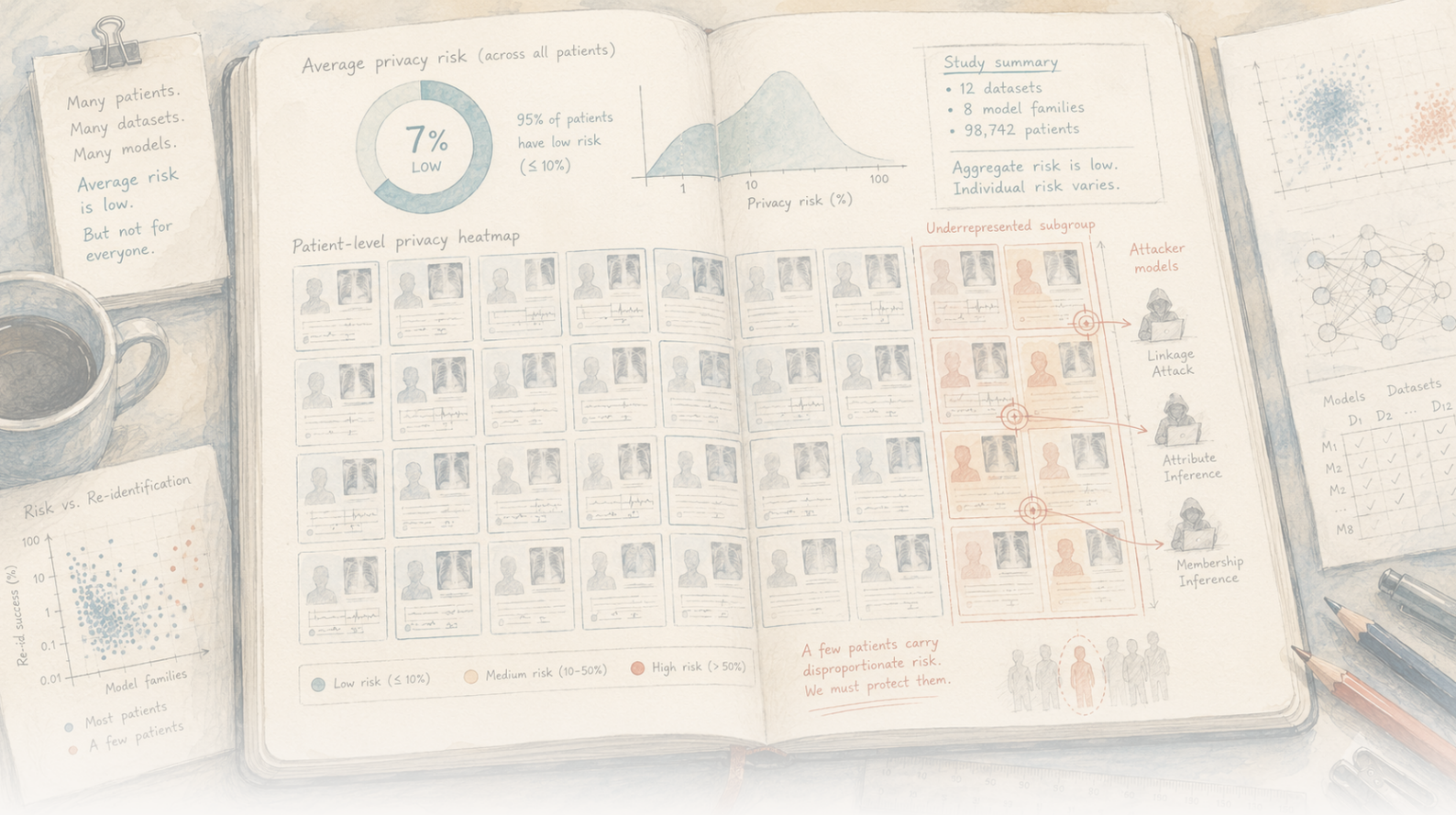


The CLEAN PAPER

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 · 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

隱私保證是對一個人的承諾，而不是對一個平均值的承諾

當一個醫療 AI 模型被稱為「隱私保護」時，這一說法通常建立在一個數字之上：在所有為訓練該模型貢獻了數據的患者當中，任何個人的成員身份能被推斷出來的平均機率很低。這聽起來令人安心。而這篇論文那個安靜而令人不安的觀點是：這是一個錯誤的數字。

隱私不是一個平均值。它是對每一個個體作出的承諾——即身處訓練集之中不會反過來暴露他們。而一個平均值可以對幾乎所有人都信守這一承諾，卻對少數幾個人徹底違背它。研究者們以前所未有的解析度、逐患者地衡量隱私風險，並恰恰發現了這一點：那些在匯總層面看起來很安全的模型，可能會近乎完美地洩露特定個體的成員身份——而且被它們洩露的那些個體，恰恰絕大多數是本就最缺乏保護的人。

作者們做了什麼

這個團隊——由 Moritz Knolle 領銜，成員包括 Daniel Rückert（兩人都在 Technical University of Munich）和 Georg Kaissis（Hasso Plattner Institute），以及 Imperial College London 的同事們——拿起了一種被稱為**成員推斷攻擊**的著名隱私威脅，並改變了所提的問題。他們沒有問「平均而言，這種攻擊在整個數據集上成功的頻率有多高？」，而是問「對於這一位特定的患者，他們暴露的程度有多深？」

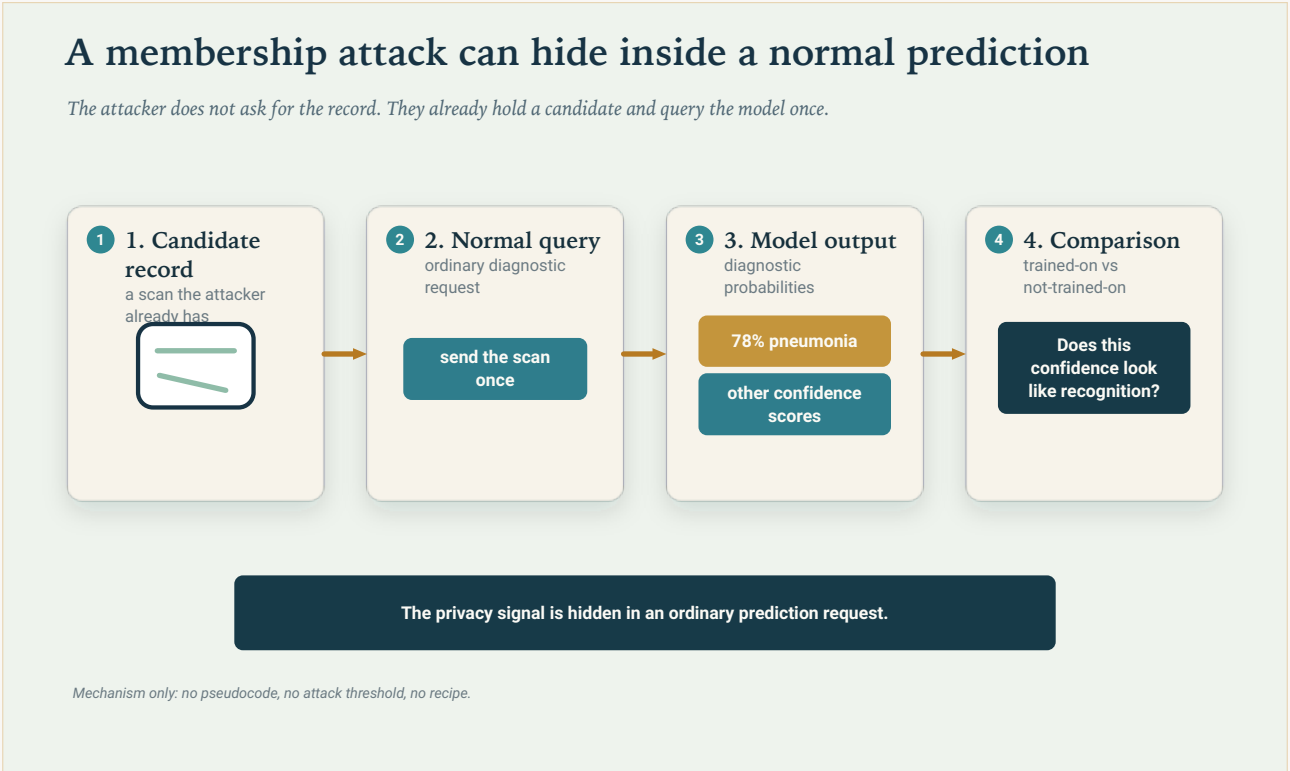
他們在**七個成熟的醫療數據集**上運行了這項逐患者分析，這些數據集橫跨了差異極大的數據類型——醫學影像、心電圖（ECG）以及電子健康記錄——並且，在這些數據集上，每個數據集各訓練了 200 個模型。對於每個模型中的每一位患者，他們估計了攻擊者能以多大的把握判斷出這個人的記錄是否曾是訓練數據的一部分，然後按群體拆解結果：疾病狀態、自我報告的種族、保險、性別以及影像採集方案。

> **成員推斷攻擊究竟是什麼**
>
> 這裡的洩露比「模型把你的記錄吐出來」要微妙得多。它關乎的是成員身份——僅僅是你的數據曾出現在訓練集中這一事實。
>
> 先從這類模型平常做什麼說起。你遞給它一張掃描片——比如說一張胸部 X 光片——它回遞給你一組機率：78% 的肺炎可能性，連同心臟肥大、水腫、實變的讀數。那個日常的診斷答案就是攻擊唯一會用到的東西。沒有任何奇技淫巧。
>
> 它所依賴的弱點是這樣的：一個模型對它所訓練過的那些確切樣本，通常會比對它從未見過的樣本**略微更有把握**。想像一個偷偷提前看過考卷的學生——在他練習過的那些題目上，答案來得快了那麼一點、篤定了那麼一點。從這個破綻出發，推算出某一份特定的記錄是否屬於模型研習過的樣本之一，這就是「成員推斷」的含義。
>
> 那麼這就是這場攻擊，一步一步來。有人想知道某個特定的人的掃描片是否被用於訓練模型。他們**並不**向模型索要那份記錄——他們手裡早已握有一張候選掃描片（那個人本人的，或一份接近的副本）。他們把它作為一次普通的診斷請求送進去一次，並記錄下答案有多篤定。然後他們檢查這份篤定看起來更像一個訓練過這張掃描片的模型，還是一個沒訓練過的模型——這

是一個他們可以廉價地做出的比較：只需在一台普通電腦上、無需任何特殊硬體，訓練一個自己的替身（一個「參考模型」），來學習那種洩露天機的過度自信長什麼樣。如果真實的模型對這張掃描片異乎尋常地確信——彷彿認得它——那就是這張掃描片曾在訓練數據中的有力證據。令人不安之處在於，這次請求沒有任何地方看起來像一場攻擊：它和臨床醫生會發出的診斷查詢一模一樣，而隱私訊號就藏在一個普通預測的內部。這正是為何這種風險是具體實在的——儘管到目前為止，它是在明確陳述的實驗室假設下被演示出來的，而非在現實世界中被逮個正著。

如果記錄本身從不洩露，成員身份為什麼重要？因為成員身份本身就是一個事實。如果一個模型是在接受過某種特定癌症免疫療法的患者身上訓練出來的，那麼確認你的記錄在其中，就揭示了你很可能患過那種癌症——這正是保險公司或雇主永遠不該能夠推斷出來的那類事情。內容始終被封存；逃逸出來的是「屬於其中」這一事實。

作者們明明白白地列出了這些假設——能訪問模型的普通預測、一份候選記錄，以及攻擊者自己的參考模型——這並非作為一份操作指南，而是為了讓這一威脅能被推理審視，而非被含糊搪塞。



這場攻擊不需要一個特殊的隱私 API。如果模型對一份它此前見過的記錄異乎尋常地自信，一次正常的診斷查詢就可能洩露出一個成員身份訊號。

他們發現了什麼

三項發現，一項比一項更為尖銳。

平均值掩蓋了那些被暴露的人。以匯總方式衡量——也就是慣常的做法——這些模型中的許多看起來私密得令人安心：對大多數患者而言，攻擊的表現並不比隨機猜測更好。而逐患者的視角講述了一個不同的故事。正如 Knolle 在這項研究的公告中所說的那樣，以往的評估「始終只衡量了所有患者的平均風險。我們首次在個體患者的層面上審視了風險——它描繪出一幅截然不同的圖景。」對於某些個體，攻擊成功率**近乎完美**——作者們將其衡量為一個 **0.95 或更高**的攻擊 AUC（0.5 是擲硬幣的水平，1.0 是毫無瑕疵）——而與此同時，全數據集範圍內的平均值看起來並不比隨機猜測更好。

平均值可以停留在接近隨機猜測的水平，而一小撮尾部的記錄仍然要暴露得多。這篇論文的觀點是：隱私必須在患者層面而不僅僅是在匯總層面接受檢驗。

這種暴露是不均等的——而且它落在了代表性不足的群體身上。最容易遭受近乎完美攻擊的患者，系統性地都是那些處於在**數據中代表性不足**的群體中的人：少數族裔、罕見病表型、異常的影像特徵。在電子記錄數據集中，Black patients 在最脆弱的那批記錄裡出現的頻率比預期高出 **31%**；在乳腺 X 線攝影數據集中，被標記為疑似惡性的掃描片在那個危險地帶裡過度代表的幅度驚人，達到了 **+1,179%**。（這兩個數字是例子，而非全貌——論文報告了在若干群體中同樣的偏斜——而且它們是那個極小的極端風險尾部之內的相對過度代表：指的是這些患者在最被暴露的記錄中出現的頻率高出多少，而不是 Black patients 或疑似乳腺攝影患者中被暴露的那一比例。）一個模型能用來把這些患者模糊融入其中的相似樣本更少，所以他們的記錄會凸顯出來——而凸顯出來恰恰是成員推斷攻擊所要探測的東西。這種隱私失效並非隨機；它集中落在那些本就處於邊緣的人身上。

更大的模型讓情況更糟。暴露於近乎完美攻擊之下的患者數量，隨著模型容量的增大而急劇上

升——在一個皮膚病學數據集中，這一比例從最小模型中的基本為零，攀升到了最大模型中的大約**十分之一**。隨著醫療 AI 模型變得越來越大、越來越強——這正是整個領域正在前進的方向——作者們警告說，這個特定的風險會變得更嚴重，而非更輕。

Rückert 的總結毫不含糊：「這不是一個可以容忍的風險。健康數據高度敏感。」

為什麼「平均而言安全」是錯誤的檢驗標準

人們很容易把一個較低的平均風險讀作一張健康證明。這篇論文的核心教訓是：這裡的求平均不僅僅是不夠精確——它衡量的根本就是錯誤的東西。

一項隱私保證只有在它對風險最大的那個人也成立時才有意義，而不是對處於中間位置的那個人成立就夠了。一個 1000 名患者中有 999 名無法被還原、但有 1 名能被近乎完美地識別出來的模型，在任何對那 1 個人有意義的層面上都不是「99.9% 私密」——而如果那 1 個人可以預料地就是罕見病患者或少數族裔患者，那麼這個指標就不只是不完整，它是在悄悄地施行歧視。它報告的是多數人的安全，卻把它稱作所有人的安全。

這就是這篇論文所迫使的轉變：從這個模型平均而言有多私密？轉向誰是最被暴露的患者，而他們又是誰？這是兩個不同的問題，而只有第二個才是一個隱私問題。

這項研究沒有證明——或聲稱——什麼

- 它**沒有**說醫療 AI 應該被拋棄。作者們的立場是緩解，而非撤退：衡量並修復這個風險，而不是停止建設。
- 它**並不**意味著每一個醫療模型都在洩露，或者任何一個已部署的模型已經遭到攻擊。它表明的是風險確實存在、且分佈不均，以及標準的匯總指標會漏掉它。
- 它**並不**意味著你的記錄已經被暴露了。這場攻擊需要特定的條件——能訪問模型、一份候選記錄，以及攻擊者自己的基礎設施——而不是一種隨手可得的能力。
- 它**並沒有**表明患者的內容會洩露。洩露出來的是成員身份——被納入其中這一事實——它之所以危險是出於一個不同的原因，而不是因為記錄被傾倒了出來。
- 它**並沒有**把這些差距歸結為一個單一的醒目數字。那個強有力、可復現的結果是這個模式——匯總指標低估了個體風險，而代表性不足的患者承受了其中的大部分——它橫跨多個數據集和數百個模型。

證據有多強？

這是一個實證的、方法學上的結果，而且是一個穩健的結果。這個模式——匯總隱私指標系統性地低估了逐患者風險，殘餘風險集中在代表性不足的群體身上，且它隨模型容量而惡化——在**七個不同數據類型的數據集**和每個數據集大量的模型上都成立。正是這種廣度，讓一個測量層面的論斷變得可

信，而不是淪為單一數據集的假象。

兩點誠實的告誡。第一，這是對風險的一次演示，是通過運行作者們自己構建的攻擊來衡量的；它刻畫的是一個有能力的攻擊者在既定假設下能夠做到多好，而不是現實世界中的攻擊實際發生得有多頻繁。第二，那些鮮明的數字——某些患者近乎完美的攻擊成功率——按其設計描述的就是境況最糟的那批個體；這正是全部的用意所在，它們應被讀作「尾部遠比平均值所暗示的沉重得多」，而不是「大多數患者都被暴露了」。

恰當的立場既不是驚慌也不是不屑：一項審慎的、多數據集的研究表明，我們認證醫療 AI 隱私的標準做法對它自己最糟糕的那些情形是盲目的——而那些最糟糕的情形，恰恰落在了最沒有餘地可供騰挪的患者身上。

為什麼這很重要

有兩件事讓這不止是一個技術腳註。

第一，它改變了「隱私保護」這個說法應該被允許指代什麼。如果一個模型帶著一個平均隱私分數被發布出來，那個分數可以是貨真價實地低，卻依然掩藏了一部分能被成員推斷攻擊近乎完美地識別出來的患者。這篇論文的實際訴求是具體的：在發布之前**逐患者地**評估隱私風險，控制誰能訪問已部署的模型，並使用諸如**差分隱私**這樣的技術——在訓練過程中加入的、經過小心校準的少量噪聲，它能鈍化成員推斷攻擊，代價是模型的有用性會有一份真實且受控的損失（隱私與效用之間的權衡是明擺著的，而非免費的）。「我們檢查了平均值」應該不再算作已經檢查過了。

第二，它把隱私和公平交織在了一起。那些在醫療數據中代表性不足的群體——因此本就被醫療 AI 服務得更差——結果恰恰就是他們的隱私被那個 AI 保護得最少的群體。一個正在為第一重不平等而努力的領域，不能把第二重當作別人部門的事。他們是同一批人。

那個關於醫療 AI 隱私的令人安心的故事——我們衡量過了，平均值很低，我們沒問題——正是這篇論文要拆解的故事。這樣做不是為了把任何人嚇離這項技術，而是為了把標準挪到它本該在的地方：一個只有當你已經為最可能受到傷害的那個人檢驗過它之後，你才能聲稱自己信守的承諾。

乾淨的總結

成員推斷攻擊試圖判定某個特定的人的記錄是否曾在一個 AI 模型的訓練數據之中——而由於成員身份本身可能就帶有揭示性（比如說，揭示你患過某種特定的疾病），所以即便底層的記錄從未浮出水面，這也是一次貨真價實的隱私洩露。此前的工作衡量的是這類攻擊平均而言在一個數據集上成功的頻率。這項研究衡量的是它逐患者的情況，橫跨七個醫療數據集（影像、ECG、電子記錄）以及每個數據集的許多模型，並發現匯總指標嚴重低估了風險：某些個體能被近乎完美地識別出來（攻擊 AUC $\geq$ 0.95），即便全數據集範圍的平均值看起來很安全；最脆弱的那些人系統性地來自代表性不足的群體（少數族裔、罕見病、異常影像）；而且這個問題隨模型規模的增大而增長。作者們並不主張拋棄醫療 AI——他們主張在個體層面衡量隱私、控制模型訪問權限，並使用差分隱私。要點在於：「平均而言私密」並不是一項隱私保證，而它辜負的那些人，通常就是本就最缺乏保護的人。

無廢話核查

這篇論文表明了什麼：橫跨七個醫療數據集以及每個數據集的許多模型，逐患者的成員推斷風險對某些個體而言遠高於匯總指標所暗示的水平——高達近乎完美的攻擊成功率（AUC $\geq$ 0.95）——而這份殘餘風險不成比例地落在代表性不足的群體身上，並隨模型容量的增大而增長。

什麼是貌似合理但未經證實的：現實世界中的攻擊者已經在針對已部署的臨床模型利用這一點；這項研究演示的是既定假設下的能力及其分佈，而不是現實世界中攻擊的發生頻率。

它沒有表明什麼：醫療 AI 應該被拋棄；每個模型都在洩露，或任何一個特定的已部署模型已被攻破；患者記錄的內容（而非成員身份）被暴露；一個單一的量化差距數字——那個穩健的結果是這個模式，而不是某一個數字。

主要的侷限：它是通過作者們自己的攻擊來衡量最壞情形風險的，而不是通過觀測到的事件；那些戲劇性的數字按其設計描述的是最被暴露的那批個體；而各群體之間確切的差距，是作為一個橫跨多個數據集的一致模式呈現出來的，而非被歸結為一個數字。

一位普通讀者應該抱有多大的信心？對以下幾點信心很高：匯總隱私指標低估了個體風險，殘餘風險集中在代表性不足的患者身上，以及這一點隨模型規模的增大而惡化——它是在許多數據集和模型上被演示出來的。對這類攻擊在現實世界中的發生頻率則只有中等信心，因為這項研究並沒有衡量它。穩妥的讀法是：既不是「醫療 AI 會洩露你的數據」，也不是「隱私問題已經解決」，而是「標準的隱私核查對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——所以這套核查必須改變。」

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-ai

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