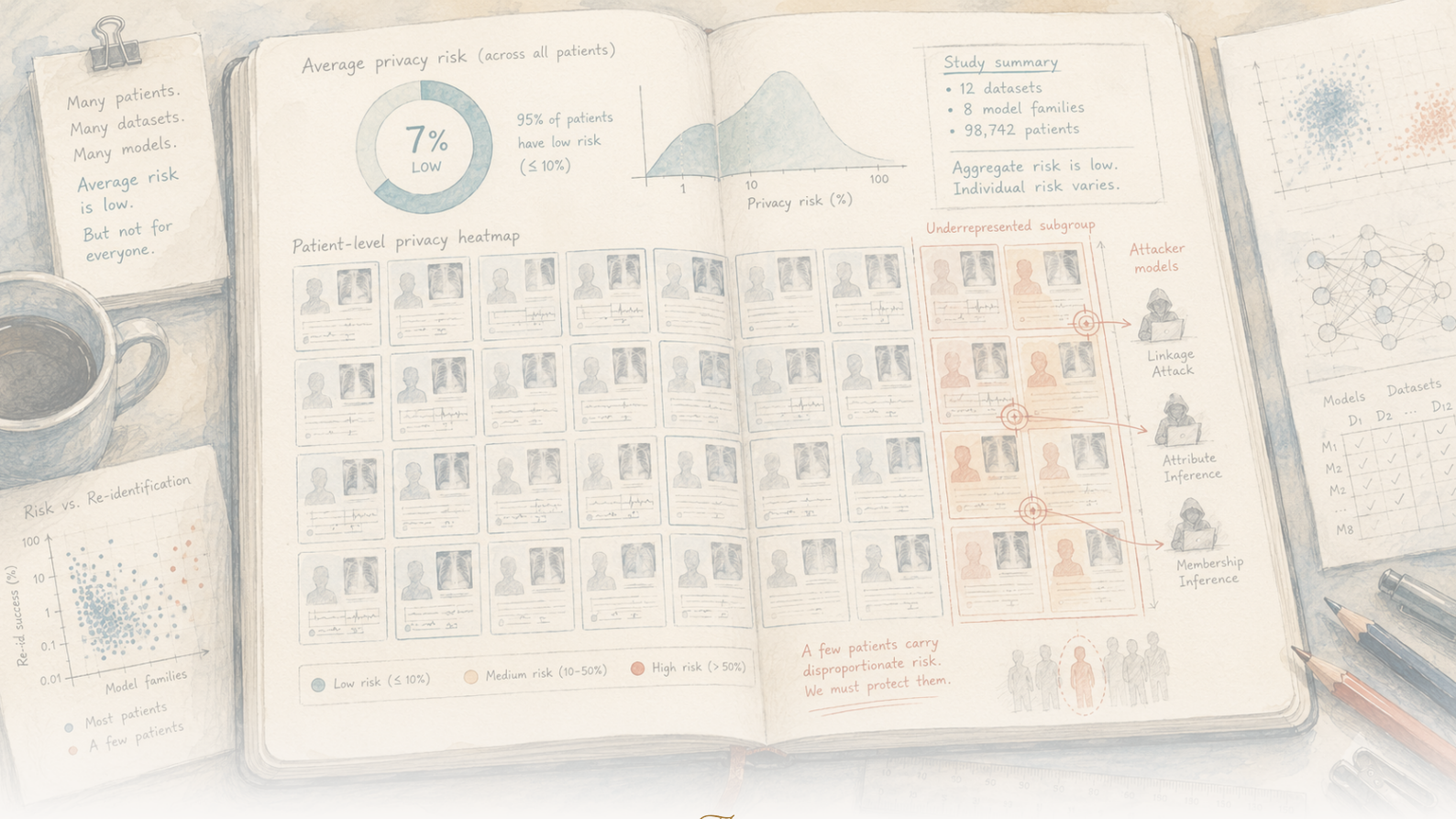


The Clean Paper

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

یک AI پزشکی می‌تواند در میانگین خصوصی باشد و باز هم بعضی بیماران را افشا کند — بیش از همه کم‌نماینده‌ها را

مدل *AI* پزشکی که «حافظ حریم خصوصی» نامیده می‌شود معمولاً بر یک عدد میانگین تکیه دارد. این مطالعه می گوید آزمون اشتباه همین است. با بررسی *attack inference membership* — که می‌سنجد آیا رکورد یک فرد مشخص در دادهٔ آموزش بوده و بنابراین می‌تواند حتی بدون افشای خود پرونده اطلاعات بیماری را لو دهد — نویسندگان خطر را به‌جای *aggregate* برای هر بیمار، در هفت *dataset* پزشکی و مدل‌های متعدد اندازه گرفتند. الگو: مدل‌هایی که در میانگین امن‌اند می‌توانند افراد خاصی را تقریباً بی‌نقص قابل‌شناسایی کنند ($AUC \geq 95.0$)؛ افراد در معرض به‌طور نظام‌مند از گروه‌های کم‌نماینده‌اند؛ و خطر با بزرگ شدن مدل افزایش می‌یابد. توصیه ترک *AI* پزشکی نیست، بلکه سنجش حریم خصوصی در سطح بیمار، کنترل دسترسی و *privacy dfferential* است. هستهٔ ناراحت کننده: «در میانگین خصوصی» تضمین حریم خصوصی نیست و برای همان بیمارانی شکست می‌خورد که از قبل کمتر محافظت شده‌اند.

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: 0-10688-026-s41586/1038.10

تضمین حریم خصوصی وعده‌ای به یک فرد است، نه به یک میانگین

وقتی یک مدل AI پزشکی «حافظ حریم خصوصی» نامیده می‌شود، این ادعا معمولاً بر یک عدد تکیه دارد: در میان تمام بیمارانی که داده‌شان مدل را آموزش داده، به‌طور متوسط احتمال اینکه عضویت هر فرد در مجموعهٔ آموزش قابل‌استنباط باشد پایین است. آرامش‌بخش به نظر می‌رسد. نکتهٔ آرام و ناراحت‌کنندهٔ این مقاله این است که این عدد اشتباه است.

حریم خصوصی میانگین نیست. وعده‌ای است به هر فرد — اینکه بودن در مجموعهٔ آموزش بعداً به افشای او برنگردد. و یک میانگین می‌تواند این وعده را برای تقریباً همه نگه دارد و برای چند نفر کاملاً بشکند. پژوهشگران خطر حریم خصوصی را بیماربه‌بیمار، با وضوحی که پیش از این استفاده نشده بود، اندازه گرفتند و دقیقاً همین را یافتند: مدل‌هایی که در مجموع امن به نظر می‌رسند می‌توانند عضویت بعضی افراد خاص را تقریباً بی‌نقص فاش کنند — و افرادی که فاش می‌شوند به‌طور نامتناسب همان‌هایی هستند که از قبل کمترین حفاظت را دارند.

نویسندگان چه کردند

تیم — به رهبری Moritz Knolle، Rückert Daniel همراه (هر دو از Technical University of Munich) و Georg Kaissis، Hasso Plattner (Institute) با همکارانی از Imperial College London — یک تهدید شناخته‌شدهٔ حریم خصوصی به نام **membership inference attack** را گرفت و سؤال را عوض کرد. به‌جای اینکه بپرسد «این حمله به‌طور متوسط چند بار در dataset موفق می‌شود؟»، پرسید «این بیمار مشخص چقدر در معرض است؟»

آن‌ها این تحلیل بیماربه‌بیمار را روی **هفت dataset پزشکی تثبیت‌شده** با انواع دادهٔ بسیار متفاوت — تصویربرداری پزشکی، الکتروکاردیوگرام و پروندهٔ سلامت الکترونیکی — و در هر dataset روی ۲۰۰ مدل اجرا کردند. برای هر بیمار در هر مدل برآورد کردند مهاجم با چه اطمینانی می‌تواند بگوید رکورد آن فرد بخشی از دادهٔ آموزش بوده یا نه، سپس نتایج را بر اساس گروه شکستند: وضعیت بیماری، race خوداظهاری، بیمه، جنس و پروتکل تصویربرداری.

membership inference attack واقعاً چیست

در شما دادهٔ که واقعیت این صرف — است عضویت دربارهٔ است. می‌ریزد» بیرون را پرونده‌ات «مدل از ظریف‌تر اینجا نشت است. بوده آموزش مجموعهٔ

از کاری شروع کنید که چنین مدلی معمولاً انجام می‌دهد. مثلاً یک X-ray قفسهٔ سینه به آن می‌دهید و احتمال‌ها پس می‌گیرید: ٪۷۸ احتمال ،pneumonia همراه خوانش‌هایی برای ،cardiomegaly oedema و .consolidation حمله فقط از همین پاسخ تشخیصی روزمره استفاده می‌کند. هیچ API عجیب یا درپشتی لازم نیست.

ضعفی که استفاده می‌کند این است: مدل معمولاً دربارهٔ دقیقاً همان نمونه‌هایی که رویشان آموزش دیده کمی **مطمئن‌تر** از نمونه‌هایی است که هرگز ندیده. مثل دانش‌آموزی که مخفیانه آزمون را قبلاً دیده باشد — روی سؤال‌هایی که تمرین کرده، پاسخ کمی سریع‌تر و با اطمینان بیشتری می‌آید. استنباط اینکه یک رکورد خاص یکی از مثال‌های آموزشی بوده یا نه از همین نشانه، همان membership» «inference است.

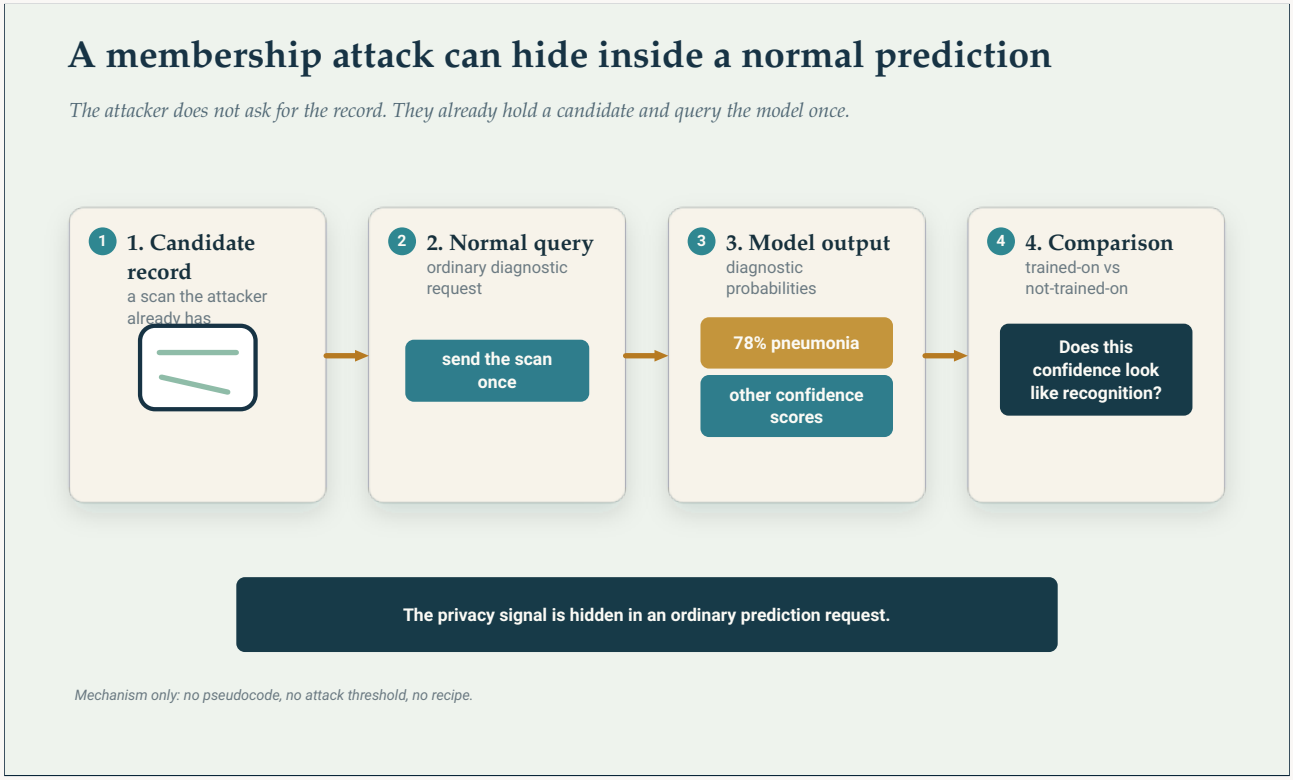
حمله قدم‌به‌قدم چنین است. کسی می‌خواهد بداند scan یک فرد مشخص برای آموزش مدل استفاده شده یا نه. از مدل پرونده را درخواست **نمی‌کند** — از قبل یک scan نامزد، مثلاً scan خود فرد یا نسخه‌ای بسیار نزدیک، دارد. آن را یک بار مثل درخواست تشخیصی معمول به مدل می‌دهد و می‌بیند مدل چقدر مطمئن است. سپس بررسی می‌کند این سطح اطمینان بیشتر شبیه مدلی است که روی این *scan* آموزش دیده یا مدلی که ندیده. برای این مقایسه می‌تواند یک مدل مرجع خودش را با هزینهٔ نسبتاً کم آموزش دهد تا الگوی آن overconfidence مشخص را یاد بگیرد، روی رایانه‌ای معمول و بدون سخت‌افزار خاص.

اگر مدل واقعی دربارهٔ این scan به‌شکل غیرعادی مطمئن باشد — انگار آن را می‌شناسد — شاهد قوی‌ای است که scan در دادهٔ آموزش بوده است.

بخش ناراحت‌کننده این است که هیچ چیز در درخواست شبیه حمله نیست: همان query تشخیصی است که یک clinician می‌فرستد و سیگنال حریم خصوصی داخل prediction عادی پنهان است. دقیقاً به همین دلیل خطر ملموس است — هرچند تا اینجا زیر فرض‌های آزمایشگاهی مشخص نشان داده شده، نه اینکه در دنیای واقعی act the in caught شده باشد.

چرا عضویت مهم است اگر خود رکورد هرگز نشت نکند؟ چون عضویت خودش یک واقعیت است. اگر مدلی روی بیمارانی آموزش دیده که یک immunotherapy خاص سرطان دریافت کرده‌اند، تأیید اینکه رکورد شما داخل آن بوده نشان می‌دهد احتمالاً همان سرطان را داشته‌اید — چیزی که بیمه‌گر یا کارفرما نباید بتواند استنباط کند. محتوای پرونده مهروموم می‌ماند؛ واقعیت تعلق است که بیرون می‌آید.

نویسندگان این فرض‌ها را روشن می‌گویند — دسترسی به هایprediction عادی مدل، یک رکورد نامزد و مدل مرجع خود مهاجم — نه به‌عنوان how-to، بلکه برای اینکه تهدید را بتوان منطقی بررسی کرد نه اینکه با دست تکان دادن کنار گذاشت.



حمله به API ویژهٔ حریم خصوصی نیاز ندارد. اگر مدل روی رکوردی که قبلاً دیده به‌طور غیرعادی مطمئن باشد، یک query تشخیصی معمول می‌تواند سیگنال عضویت نشت دهد.

Original Aurora diagram — The Clean Paper CC BY 0.4

چه پیدا کردند

سه یافته، هر کدام تیزتر از قبلی.

میانگین‌ها افراد در معرض را پنهان می‌کنند. وقتی به شکل aggregate — روش معمول — اندازه گرفته شود، بسیاری از این مدل‌ها آرامش‌بخش و خصوصی به نظر می‌رسند: حمله برای بیشتر بیماران بهتر از شانس نیست. دید بیماربه‌بیمار داستان دیگری گفت. همان‌طور که Knolle در اعلام مطالعه گفت، ارزیابی‌های قبلی «فقط خطر میانگین میان همهٔ بیماران را اندازه گرفته‌اند. ما برای نخستین بار خطر را در سطح هر بیمار بررسی کردیم — و تصویر بسیار متفاوتی می‌دهد.» برای بعضی افراد، حمله **تقریباً بی‌نقص** موفق می‌شود — نویسندگان آن را AUC attack **95.0 یا بالاتر** اندازه می‌گیرند (5.0 مثل شیر یا خط، 0.1 بی‌نقص) — حتی وقتی میانگین کل dataset بهتر از شانس به نظر نمی‌رسید.



میانگین می‌تواند نزدیک شانس بماند، در حالی که دنباله‌ای کوچک از رکوردها بسیار بیشتر در معرض‌اند. نکتهٔ مقاله این است که حریم خصوصی باید در سطح بیمار بررسی شود، نه فقط aggregate.

Original Aurora diagram — The Clean Paper CC BY 0.4

مواجهه نابرابر است — و روی کم‌نماینده‌ها می‌افتد. بیمارانی که در برابر حملهٔ تقریباً بی‌نقص آسیب‌پذیرتر بودند به‌طور نظام‌مند از گروه‌هایی می‌آمدند که در داده **کم‌نماینده** بودند: اقلیت‌های قومی، هایphenotype بیماری نادر، ویژگی‌های تصویربرداری نامعمول. در dataset پروندهٔ الکترونیکی، بیماران Black **۳۱٪ بیشتر از انتظار** میان رکوردهای بسیار آسیب‌پذیر دیده شدند، در مجموعهٔ mammography، هایscan علامت گذاری‌شده به‌عنوان مشکوک به بدخیمی در آن ناحیهٔ خطر با **+۱۱۷۹٪** overrepresentation حضور داشتند. (این دو عدد مثال‌اند، نه کل تصویر — مقاله skew مشابهی در چند گروه گزارش می‌کند — و *overrepresentation* نسبی در دنبالهٔ کوچک با خطر بسیار بالا هستند: یعنی این بیماران در میان رکوردهای بسیار در معرض چقدر بیش از انتظار دیده می‌شوند، نه اینکه چه درصدی از همهٔ بیماران Black یا mammography مشکوک «افشا شده‌اند».) مدل مثال‌های مشابه کمتری دارد تا این بیماران در میانشان محو شوند، بنابراین رکوردشان برجسته می‌ماند — و برجسته بودن دقیقاً چیزی است که attack membership تشخیص می‌دهد. شکست حریم خصوصی تصادفی نیست؛ روی کسانی متمرکز می‌شود که از قبل در حاشیه‌اند.

مدل‌های بزرگ‌تر بدترش می‌کنند. تعداد بیمارانی که در معرض حملهٔ تقریباً بی‌نقص بودند با ظرفیت مدل به‌شدت افزایش یافت — در یک dataset پوست، سهم از تقریباً صفر در کوچک‌ترین مدل به حدود **یک نفر از هر ده نفر** در بزرگ‌ترین مدل رسید. نویسندگان هشدار می‌دهند با بزرگ‌تر و توانمندتر شدن مدل‌های AI پزشکی — جهت حرکت کل حوزه — این خطر مشخص شدیدتر می‌شود، نه کمتر.

خلاصهٔ Rückert صریح است: «این خطر قابل‌تحمل نیست. دادهٔ سلامت بسیار حساس است.»

چرا «در میانگین امن» آزمون اشتباه است

وسوسه این است که خطر میانگین پایین را تأییدیهٔ سلامت بدانیم. درس مرکزی مقاله این است که averaging اینجا صرفاً کم‌دقت نیست — چیز اشتباهی را اندازه می‌گیرد.

تضمین حریم خصوصی فقط وقتی معنا دارد که برای فردی که بیشترین خطر را دارد برقرار باشد، نه فرد وسط توزیع. مدلی که در آن ۹۹۹ نفر از ۱۰۰۰ نفر قابل‌شناسایی نیستند اما یک نفر تقریباً بی‌نقص شناسایی می‌شود، در هیچ معنایی که برای آن یک نفر مهم باشد «۹۹٫۹٪ خصوصی» نیست — و اگر آن یک نفر به‌طور قابل‌پیش‌بینی بیمار بیماری نادر یا عضو اقلیت قومی باشد، metric فقط ناقص نیست، بی‌سروصدا تبعیض‌آمیز هم هست. امنیت اکثریت را گزارش می‌کند و نامش را امنیت همه می‌گذارد.

تغییری که مقاله تحمیل می‌کند همین است: از این مدل به‌طور متوسط چقدر خصوصی است؟ به در معرض‌ترین بیمار کیست، و چه کسی است؟ این‌ها دو سؤال متفاوت‌اند و فقط دومی سؤال حریم خصوصی واقعی است.

این مقاله چه چیزی را اثبات یا ادعا نمی کند

- **نمی‌گوید** AI پزشکی باید رها شود. چارچوب نویسندگان mitigation است، نه عقب‌نشینی: خطر را اندازه بگیرید و اصلاح کنید، ساختن را متوقف نکنید.
- **معنی‌اش این نیست** که هر مدل پزشکی نشت دارد یا هر مدل شده‌ایdeploy مورد حمله قرار گرفته. نشان می‌دهد خطر وجود دارد و نابرابر توزیع شده و هایaggregate metric معمول آن را از دست می‌دهند.
- **معنی‌اش این نیست** که پروندهٔ شما همین حالا افشا شده. حمله شرایط خاص می‌خواهد — دسترسی به مدل، رکورد نامزد و زیرساخت خود مهاجم — نه توانایی‌ای casual.
- **نشان نمی‌دهد محتوای پروندهٔ بیمار نشت می‌کند.** چیزی که نشت می‌کند عضویت است — واقعیت حضور — که به دلیل دیگری خطرناک است، نه چون کل record بیرون ریخته شود.
- هاdisparity را به یک عدد تیترگونه **تقلیل نمی‌دهد.** نتیجهٔ قوی و تکرارشونده الگو است — هایaggregate metric خطر فردی را کم‌تر نشان می‌دهند و بیماران کم‌نماینده سهم بیشتری از آن دارند — در هاdataset و صدها مدل.

قدرت شواهد چقدر است؟

این نتیجه‌ای تجربی و روش‌شناختی و مقاوم است. الگو — اینکه هایmetric حریم خصوصی aggregate به‌طور نظام‌مند خطر بیماربه‌بیمار را کمتر از واقع نشان می‌دهند، خطر باقی‌مانده روی گروه‌های کم‌نماینده متمرکز می‌شود و با ظرفیت مدل بدتر می‌شود — در **هفت dataset با انواع دادهٔ متفاوت** و تعداد زیادی مدل در هر dataset پابرجا بود. همین گستردگی است که ادعای اندازه‌گیری را معتبر می‌کند، نه artfiact یک dataset.

دو caveat صادقانه. اول، این نمایش خطر است که با اجرای حمله‌های ساخته‌شده توسط خود نویسندگان اندازه گرفته شده؛ مشخص می‌کند یک مهاجم توانمند زیر فرض‌های گفته‌شده چه می‌تواند بکند، نه اینکه حملهٔ واقعی چند بار رخ می‌دهد. دوم، عددهای چشمگیر — موفقیت نزدیک به کامل برای بعضی بیماران — به‌عمد دربارهٔ بدترین افراد توزیع‌اند؛ تمام نکته همین است و باید به شکل «دنباله بسیار

سنگین‌تر از چیزی است که میانگین نشان می‌دهد» خوانده شود، نه «بیشتر بیماران افشا شده‌اند».

موضع درست نه alarm است نه :dismissal مطالعه‌ای دقیق و چند-dataset که نشان می‌دهد روش استاندارد صدور گواهی حریم خصوصی AI پزشکی نسبت به بدترین موردهای خودش کور است — و این بدترین‌ها روی بیمارانی می‌افتند که کمترین حاشیهٔ امن را دارند.

چرا مهم است

دو چیز این را فراتر از پاورقی فنی می‌برد.

اول، معنای مجاز «حافظ حریم خصوصی» را عوض می‌کند. اگر مدل با یک امتیاز میانگین حریم خصوصی منتشر شود، آن امتیاز می‌تواند واقعاً پایین باشد و در عین حال یsubset از بیماران را پنهان کند که attack membership تقریباً بی‌نقص شناسایی‌شان می‌کند. خواستهٔ عملی مقاله مشخص است: پیش از انتشار، خطر حریم خصوصی را **برای هر بیمار** بسنجید، کنترل کنید چه کسی به مدل شدهdeploy دسترسی دارد و از تکنیک‌هایی مانند **privacy dffierential** استفاده کنید — نویز کوچک و دقیقاً کالیبره‌شده در آموزش که attack membership را کند می‌کند، با هزینه‌ای واقعی و مدیریت‌شده برای usefulness مدل (trade-off حریم خصوصی–utility صریح است، رایگان نیست). «میانگین را چک کردیم» دیگر نباید معادل «حریم خصوصی را چک کردیم» حساب شود.

دوم، حریم خصوصی را به fairness می‌بافد. همان گروه‌هایی که در دادهٔ پزشکی کم‌نماینده‌اند — و بنابراین AI پزشکی از قبل به آن‌ها خدمات بدتری می‌دهد — همان‌هایی‌اند که حریم خصوصی‌شان هم کمتر محافظت می‌شود. حوزه‌ای که روی نابرابری اول سخت کار می‌کند نمی‌تواند دومی را کار بخش دیگری بداند. آدم‌ها همان‌اند.

داستان آرامش‌بخش حریم خصوصی AI پزشکی — اندازه گرفتیم، میانگین پایین است، پس خوبیم — همان داستانی است که این مقاله باز می‌کند. نه برای ترساندن مردم از فناوری، بلکه برای جابه‌جا کردن استاندارد به جای درست: وعده‌ای که فقط وقتی می‌توانید ادعا کنید نگه داشته‌اید که آن را برای کسی که بیشترین احتمال آسیب دارد بررسی کرده باشید.

خلاصهٔ تمیز

attack inference membership تلاش می‌کند تشخیص دهد آیا رکورد یک فرد مشخص در دادهٔ آموزش مدل AI بوده است — و چون عضویت خودش می‌تواند اطلاعات حساسی فاش کند (مثلاً اینکه بیماری خاصی داشته‌اید)، حتی وقتی خود پرونده هرگز بیرون نمی‌آید یک نشت واقعی حریم خصوصی است. پژوهش قبلی میزان موفقیت حمله را به‌طور میانگین روی dataset می‌سنجید. این مطالعه آن را برای هر بیمار، در هفت dataset پزشکی (تصویربرداری، ECG، پروندهٔ الکترونیکی) و تعداد زیادی مدل در هر کدام اندازه گرفت و یافت هایaggregate metric به‌شدت خطر را کم نشان می‌دهند: بعضی افراد حتی وقتی میانگین dataset امن به نظر می‌رسد تقریباً بی‌نقص قابل‌شناسایی‌اند (attack AUC $\geq$ 95.0)؛ آسیب‌پذیرترین‌ها به‌طور نظام‌مند از گروه‌های کم‌نماینده‌اند (اقلیت قومی، بیماری نادر، تصویربرداری غیرمعمول)؛ و مشکل با اندازهٔ مدل رشد می‌کند. نویسندگان نمی‌گویند AI پزشکی را کنار بگذارید — می‌گویند حریم خصوصی را در سطح فرد اندازه بگیرید، دسترسی به مدل را کنترل کنید و privacy dffierential به کار ببرید. takeaway: «به‌طور میانگین خصوصی» تضمین حریم خصوصی نیست، و شکستش معمولاً روی کسانی می‌افتد که از قبل کمترین حفاظت را دارند.

بررسی بی‌پرده

آنچه مقاله نشان می‌دهد: در هفت dataset پزشکی و بسیاری مدل در هر کدام، خطر inference membership برای بعضی بیماران بسیار بیشتر از چیزی است که هایaggregate metric نشان می‌دهند — تا موفقیت تقریباً بی‌نقص (AUC $\geq$ 95.0) — و این خطر باقی‌مانده به‌طور نامتناسب روی گروه‌های کم‌نماینده می‌افتد و با ظرفیت مدل رشد می‌کند.

آنچه معقول است اما اثبات نشده: اینکه مهاجمان واقعی هم‌اکنون علیه مدل‌های بالینی شدهdeploy از این استفاده می‌کنند؛ مطالعه توانایی و توزیع خطر را زیر فرض‌های مشخص نشان می‌دهد، نه فراوانی حمله در دنیای واقعی.

آنچه نشان نمی‌دهد: اینکه AI پزشکی باید رها شود؛ هر مدل نشت می‌دهد یا یک مدل شدهٔdeploy مشخص breach شده؛ محتوای پروندهٔ بیمار، نه فقط عضویت، افشا می‌شود؛ یا یک عدد disparity واحد تمام مسئله را خلاصه می‌کند — نتیجهٔ مقاوم خود الگو است.

محدودیت‌های اصلی: risk worst-case را با حمله‌های خود نویسندگان اندازه می‌گیرد نه هایincident مشاهده‌شده؛ عددهای dramatic به‌عمد بدترین افراد را توصیف می‌کنند؛ و هایdisparity دقیق هر گروه به شکل الگویی سازگار در هاdataset نشان داده می‌شوند نه یک عدد واحد.

یک خوانندهٔ عمومی چقدر باید مطمئن باشد؟ اطمینان بالا که هایaggregate metric خطر فردی را کمتر نشان می‌دهند، خطر باقی‌مانده روی بیماران کم‌نماینده متمرکز است و با اندازهٔ مدل بدتر می‌شود — روی هاdataset و مدل‌های متعدد نشان داده شده است. اطمینان متوسط دربارهٔ فراوانی حملهٔ واقعی، که مطالعه اندازه نمی‌گیرد. خوانش امن: نه «پزشکی AI داده‌ات را نشت می‌دهد»، نه «حریم خصوصی حل شده»، بلکه «چک استاندارد حریم خصوصی نسبت به بدترین موردهای خودش کور است و آن موردها روی آسیب‌پذیرترین بیماران می‌افتند — پس خود چک باید عوض شود.»

منابع

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

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

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