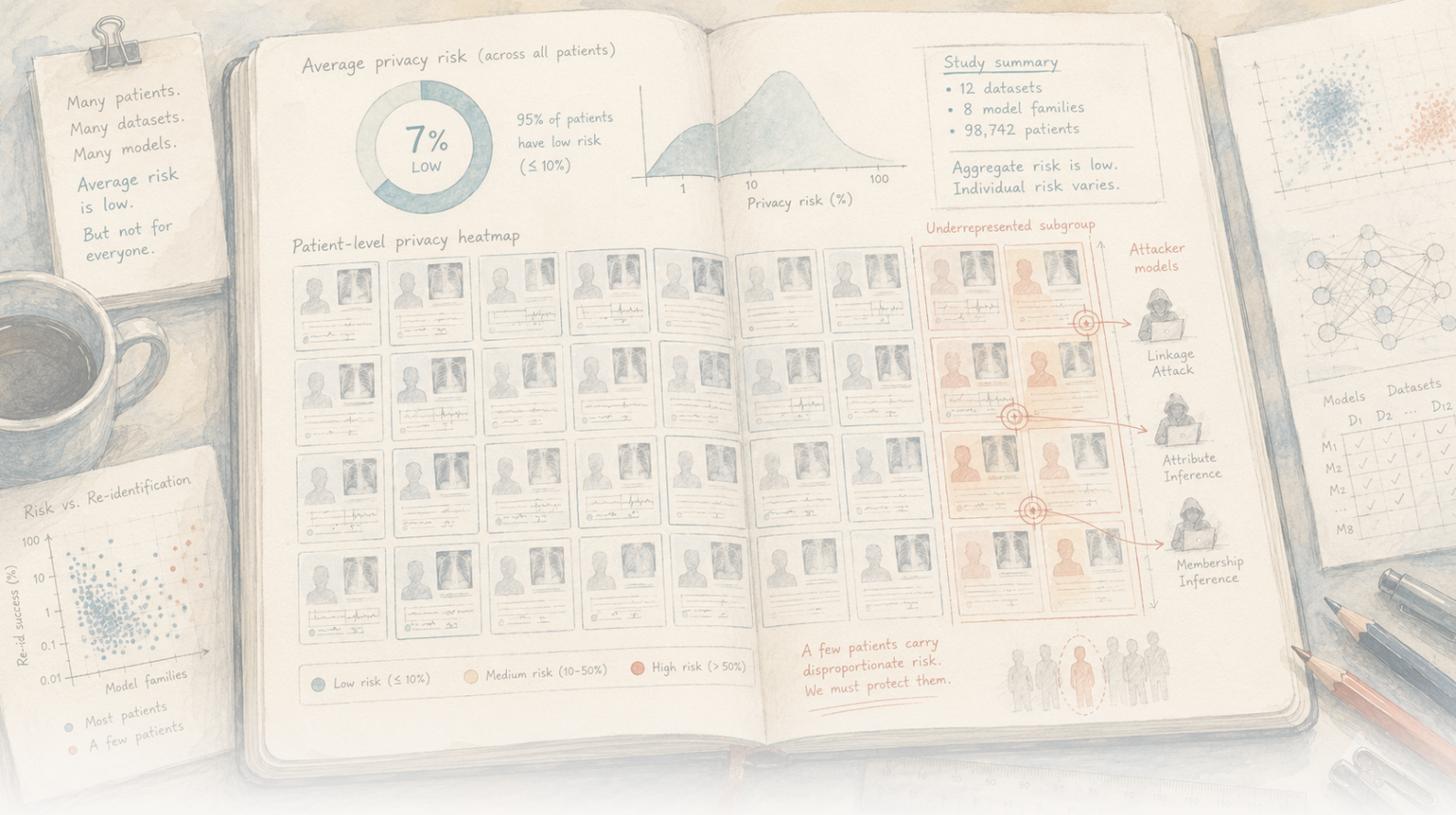


The CLEAN PAPER

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

قد يكون الذكاء الاصطناعي الطبي خاصًا في المتوسط ومع ذلك يكشف مرضى بعينهم — وناقصي التمثيل أكثر من غيرهم

عادة ما يستند وصف نموذج ذكاء اصطناعي طبي بأنه «يحافظ على الخصوصية» إلى رقم متوسط واحد. تجادل هذه الدراسة بأن ذلك هو الاختبار الخطأ. عبر هجمات استدلال العضوية — التي تكشف ما إذا كان سجل شخص بعينه ضمن بيانات تدريب النموذج، وبالتالي قد تكشف أنه كان مصاباً بمرض معين — قاس الباحثون الخطر لكل مريض بدلًا من المتوسط، عبر سبع مجموعات بيانات طبية (تصوير، *ECG*، سجلات إلكترونية) ونماذج كثيرة لكل منها. النمط: نماذج تبدو آمنة في المتوسط قد تسمح بالتعرف على أفراد محددين تقريباً بصورة كاملة (*AUC* للهجوم ≥ 0.95)؛ والأكثر انكشافاً من المجموعات ناقصة التمثيل؛ ويزداد الخطر مع نمو النماذج. لا يقول المؤلفون اتركوا الذكاء الاصطناعي الطبي، بل قيسوا الخصوصية لكل مريض واضبطوا الوصول واستخدموا الخصوصية التفاضلية. الجوهر غير المريح: «خاص في المتوسط» ليس ضماناً، ويفشل تحديداً مع المرضى الأقل حماية أصلاً.

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

ضمان الخصوصية وعد لشخص، لا لمتوسط

عندما يُقال إن نموذجًا للذكاء الاصطناعي الطبي «يحافظ على الخصوصية»، يستند هذا الادعاء عادة إلى رقم واحد: عبر جميع المرضى الذين استُخدمت بياناتهم لتدريب النموذج، يكون متوسط احتمال استنتاج ما إذا كان سجل فرد بعينه ضمن بيانات التدريب منخفضًا. هذا يبدو مطمئنًا. لكن النقطة الهادئة وغير المريحة في هذه الورقة هي أن هذا هو الرقم الخطأ.

الخصوصية ليست متوسطًا. إنها وعد لكل فرد — بأن وجود بياناته في مجموعة التدريب لن يعود ليكشفه. ويمكن لمتوسط أن يفي بهذا الوعد تقريبًا للجميع بينما يخرقه تمامًا لقلة. قاس الباحثون خطر الخصوصية مريضًا بمريض، بدقة لم تُستخدم من قبل، ووجدوا هذا بالضبط: نماذج تبدو آمنة إجمالًا يمكن أن تسرّب عضوية أفراد محددين في مجموعة التدريب بدقة تكاد تكون كاملة — والأفراد الذين تتسرب عضويتهم أكثر من غيرهم هم، بصورة غير متناسبة، من كانوا أصلًا الأقل حماية.

ما الذي فعله الباحثون

أخذ الفريق — بقيادة Moritz Knolle، ومع Daniel Rückert (كلاهما في Technical University of Munich) وGeorg Kaissis (Imperial College London إلى جانب زملاء في Hasso Plattner Institute)، تهديدًا معروفًا للخصوصية يسمى **هجوم استدلال العضوية** (membership inference attack)، وغيّر السؤال. بدل أن يسأل «في المتوسط، كم مرة ينجح هذا الهجوم عبر مجموعة البيانات؟»، سأل «بالنسبة إلى هذا المريض تحديدًا، ما مقدار انكشافه؟».

أجروا هذا التحليل لكل مريض عبر سبع **مجموعات بيانات طبية معروفة** تغطي أنواعًا مختلفة جدًا من البيانات — تصوير طبي، وتخطيط القلب الكهربائي، وسجلات صحية إلكترونية — وعلى 200 نموذج لكل مجموعة. لكل مريض في كل نموذج، قدّروا مدى ثقة مهاجم في تقرير ما إذا كان سجل ذلك الشخص قد دخل بيانات التدريب، ثم فصلوا النتائج بحسب المجموعة: حالة المرض، والعرق المبلغ عنه ذاتيًا، والتأمين، والجنس، وبروتوكول التصوير.

ما هو هجوم استدلال العضوية فعليًا

التدريب. مجموعة ضمن كانت بياناتك أن حقيقة مجرد — بالعضوية يتعلق إنه الطبي». سجلك يطبع «النموذج أن من أدق هنا التسرب ابدأ بما يفعله نموذج من هذا النوع عادة. تعطيه صورة — أشعة صدر مثلًا — فيعيد احتمالات: 78% احتمال التهاب رئوي، مع تقديرات لتضخم القلب والوذمة والتكثف. هذه الإجابة التشخيصية اليومية هي الشيء الوحيد الذي يستخدمه الهجوم. لا شيء غريبًا أو سريًا.

نقطة الضعف التي يستغلها هي أن النموذج يكون عادة **أكثر ثقة قليلًا** في الأمثلة الدقيقة التي تدرب عليها مقارنة بأمثلة لم يرها. تخيل طالبًا رأى الامتحان سرًا قبل موعده — في الأسئلة التي تدرب عليها تأتي الإجابات بسرعة وثقة أكثر قليلًا مما ينبغي. استنتاج ما إذا كان سجل معين من الأمثلة التي درسها النموذج اعتمادًا على هذه العلامة هو ما يعنيه «استدلال العضوية».

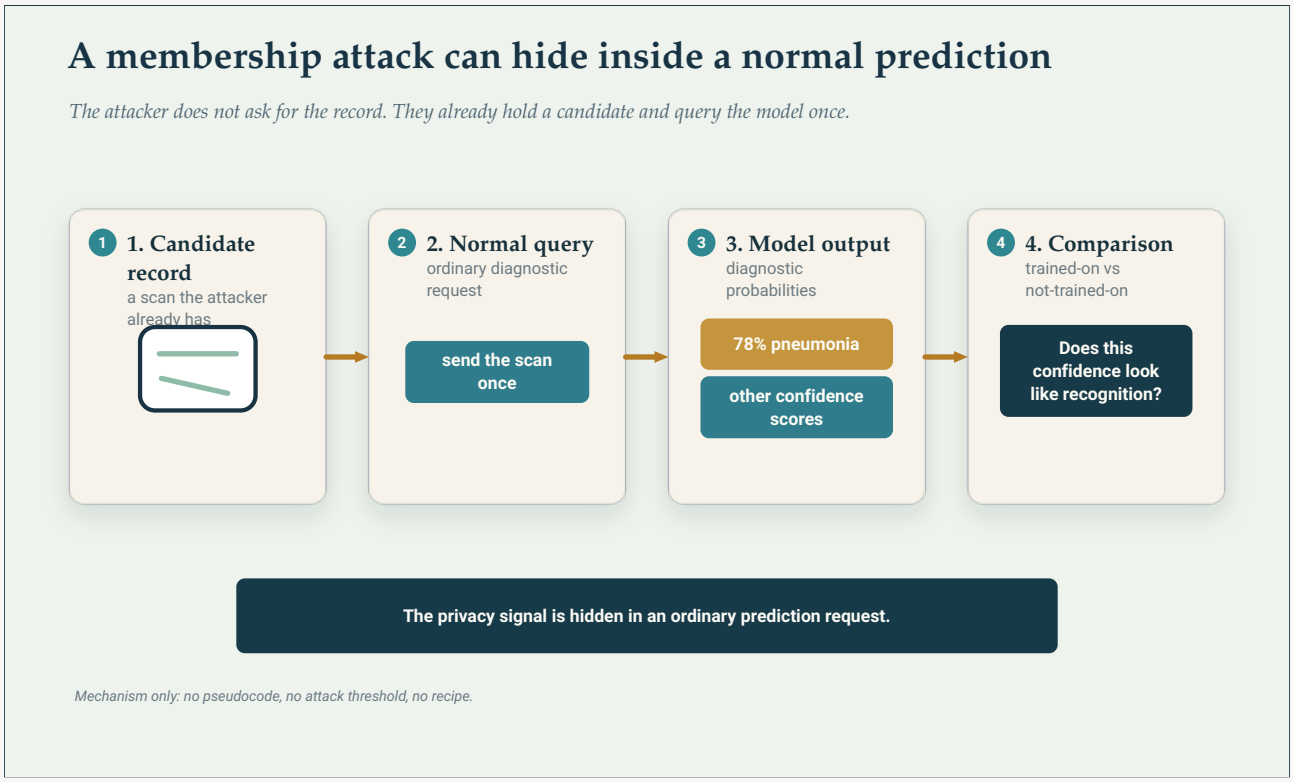
إذن الهجوم، خطوة بخطوة: يريد شخص أن يعرف هل استُخدمت صورة مريض بعينه لتدريب النموذج. هو لا يطلب من النموذج أن يعطيه السجل — بل يملك أصلًا صورة مرشحة (صورة الشخص نفسها أو نسخة قريبة منها). يرسلها مرة واحدة كطلب تشخيص عادي، ويسجل مقدار ثقة الإجابة. ثم يقارن ما إذا كانت هذه الثقة تشبه أكثر نموذجًا تدرب على الصورة أم نموذجًا لم يتدرب عليها — ويمكنه إجراء هذه المقارنة بتكلفة منخفضة عبر تدريب نموذج بديل خاص به («نموذج مرجعي») ليتعلم شكل هذه الثقة الزائدة،

على حاسوب عادي من دون عتاد خاص. إذا كان النموذج الحقيقي واثقًا بصورة غير معتادة بهذه الصورة — كما لو كان «يتعرف» عليها — فذلك دليل قوي على أنها كانت ضمن بيانات التدريب.

الجزء المقلق هو أن الطلب لا يبدو هجومًا: إنه الطلب التشخيصي نفسه الذي قد يرسله طبيب، وإشارة الخصوصية مخفية داخل تنبؤ عادي. وهذا بالضبط ما يجعل الخطر ملموسًا — حتى لو كان حتى الآن قد أُظهر ضمن افتراضات مخبرية معلنة ولم يُرصد في البرية.

لماذا تهم العضوية إذا كان السجل نفسه لا يتسرب؟ لأن العضوية حقيقة. إذا كان نموذج قد دُرب على مرضى تلقوا علاجًا مناعيًا معينًا للسرطان، فإن تأكيد أن سجلك ضمن بياناته يكشف أنك على الأرجح كنت مصابًا بذلك السرطان — وهو نوع من المعلومات لا ينبغي لشركة تأمين أو صاحب عمل أن يستنتجه. المحتوى يبقى مغلقًا؛ حقيقة الانتماء هي التي تهرب.

يعرض المؤلفون هذه الافتراضات بوضوح — الوصول إلى تنبؤات النموذج العادية، ووجود سجل مرشح، ونموذج مرجعي يملكه المهاجم — لا بوصفها دليلًا إجرائيًا، بل لكي يمكن التفكير في التهديد بجدية بدل التلويح به بصورة غامضة.



لا يحتاج الهجوم إلى واجهة خصوصية خاصة. يمكن لطلب تشخيص عادي أن يسرّب إشارة عضوية إذا كان النموذج واثقًا بصورة غير معتادة بسجل رآه من قبل.

Original Aurora diagram — The Clean Paper CC BY 0.4

ماذا وجدوا

ثلاث نتائج، كل واحدة أشد وضوحًا من السابقة.

المتوسطات تخفي الأكثر انكشافًا. عند القياس إجمالًا — وهي الطريقة المعتادة — تبدو كثير من هذه النماذج مطمئنة من ناحية الخصوصية:

بالنسبة إلى معظم المرضى لا يتجاوز الهجوم مستوى الصدفة. لكن المنظور لكل مريض روى قصة مختلفة. وكما قال Knolle في إعلان الدراسة، كانت التقييمات السابقة «تقيس دائمًا متوسط الخطر عبر جميع المرضى. نحن فحصنا الخطر على مستوى المرضى فرادى للمرة الأولى — وتظهر صورة مختلفة جدًا». بالنسبة إلى بعض الأفراد، ينجح الهجوم **تقريبًا بصورة كاملة** — يقيس المؤلفون ذلك بمساحة تحت منحنى للهجوم (AUC) تبلغ **95.0 أو أكثر** (5.0 مثل رمية عملة، و0.1 مثالي) — حتى عندما يبدو متوسط مجموعة البيانات كلها قريبًا من الصدفة.



يمكن أن يبقى المتوسط قريبًا من الصدفة بينما تظل ذيل صغير من السجلات أكثر انكشافًا بكثير. نقطة الورقة أن الخصوصية يجب أن تُفحص على مستوى المريض، لا إجمالًا فقط.

Original Aurora diagram — The Clean Paper CC BY 0.4

الانكشاف غير متساوٍ — ويقع على ناقصي التمثيل. المرضى الأكثر عرضة لهجوم شبه مثالي كانوا بصورة منهجية من مجموعات **ناقصة التمثيل في البيانات**: أقليات عرقية، وأنماط مرضية نادرة، وخصائص تصوير غير معتادة. في مجموعة السجلات الإلكترونية، ظهر المرضى السود **أكثر بنسبة 31% مما هو متوقع** بين السجلات الأكثر عرضة؛ وفي مجموعة تصوير الثدي، كانت الصور المصنفة كمشتبه بها للخباثة ممثلة في منطقة الخطر هذه بنسبة إضافية مذهلة بلغت **+1179%**. (هذان الرقمان مثالان لا الصورة كلها — تبلغ الورقة عن الانحياز نفسه عبر مجموعات عدة — وهما زيادة نسبية في التمثيل داخل الذيل الصغير شديد الخطر: أي كم مرة يظهر هؤلاء المرضى أكثر مما يُتوقع بين أكثر السجلات انكشافًا، وليس نسبة المرضى السود أو صور الثدي المشتبه بها التي انكشفت.) لدى النموذج أمثلة مشابهة أقل لتمويه هؤلاء المرضى بينها، فتبرز سجلاتهم — والبروز هو بالضبط ما يلتقطه هجوم العضوية. فشل الخصوصية ليس عشوائيًا؛ بل يتركز في الأشخاص الموجودين أصلًا على الهوامش.

النماذج الأكبر تجعل الأمر أسوأ. ارتفع عدد المرضى المعرضين لهجمات شبه مثالية بحدة مع قدرة النموذج — في إحدى مجموعات الجلدية ارتفعت الحصة من شبه صفر في أصغر نموذج إلى نحو **واحد من كل عشرة** في الأكبر. ومع نمو نماذج الذكاء الاصطناعي الطبي وقدرتها — وهو الاتجاه العام للمجال — يحذر المؤلفون من أن هذا الخطر المحدد يزداد شدة، لا العكس.

خلاصة Rückert مباشرة: «هذا ليس خطرًا مقبولًا. البيانات الصحية شديدة الحساسية».

لماذا «آمن في المتوسط» هو الاختبار الخطأ

الإغراء هو قراءة انخفاض متوسط الخطر كشهادة صحة كاملة. الدرس الأساسي في الورقة هو أن المتوسط هنا ليس مجرد قياس غير دقيق — بل يقيس الشيء الخطأ.

ضمان الخصوصية له معنى فقط إذا صمد للشخص الأكثر عرضة للخطر، لا للشخص الموجود في الوسط. نموذج لا يمكن التعرف فيه على 999 من كل 1000 مريض لكن يمكن تحديد واحد تقريبًا بصورة كاملة ليس «خاصًا بنسبة 99.9%» بأي معنى يهم ذلك الشخص — وإذا كان ذلك الشخص بصورة متوقعة مريض مرض نادر أو فردًا من أقلية عرقية، فالمقياس ليس ناقصًا فقط بل تمييزيًا بهدوء. يبلغ عن أمان الأغلبية ويسميه أمان الجميع.

هذا هو التحول الذي تفرضه الورقة: من ما مدى خصوصية هذا النموذج في المتوسط؟ إلى من هو المريض الأكثر انكشافًا، ومن يكون؟ هذان سؤالان مختلفان، والثاني فقط هو سؤال خصوصية حقيقي.

ما الذي لا تثبته الورقة أو تدعيه

- لا تقول إن الذكاء الاصطناعي الطبي يجب التخلي عنه. إطار المؤلفين هو التخفيف لا التراجع: قِس الخطر وأصلحه، ولا تتوقف عن البناء.
- لا تعني أن كل نموذج طبي يسرّب، أو أن نموذجًا منشورًا بعينه تعرض لهجوم. إنها تُظهر أن الخطر موجود وموزع بصورة غير متساوية، وأن المقاييس الإجمالية المعتادة تفوّته.
- لا تعني أن سجلاتك مكشوفة بالفعل. يحتاج الهجوم إلى شروط محددة — الوصول إلى النموذج، وسجل مرشح، وبنية يملكها المهاجم — وليس قدرة عابرة يستطيع أي شخص استخدامها بلا موارد.
- لا تُظهر تسرب محتوى سجل المريض. الذي يتسرب هو العضوية — حقيقة الإدراج — وهي خطيرة لسبب مختلف، لا لأن السجل يُطبع للخارج.
- ولا تختزل التفاوتات في رقم واحد لافت. النتيجة القوية القابلة للتكرار هي النمط: المقاييس الإجمالية تقلل خطر الفرد، والمرضى ناقصو التمثيل يتحملون معظمه، عبر مجموعات بيانات ومئات النماذج.

ما مدى قوة الدليل؟

هذه نتيجة تجريبية ومنهجية، وهي قوية. النمط — أن مقاييس الخصوصية الإجمالية تقلل بصورة منهجية الخطر لكل مريض، وأن الخطر المتبقي يتركز في المجموعات ناقصة التمثيل، وأنه يسوء مع قدرة النموذج — ظهر عبر **سبع مجموعات بيانات من أنواع مختلفة** وعدد كبير من النماذج لكل مجموعة. هذا الاتساع هو بالضبط ما يجعل ادعاء القياس موثوقًا بدل أن يكون أثرًا خاصًا بمجموعة واحدة.

هناك تحفظان صريحان. أولًا، هذه دراسة خطر تُقاس بتشغيل الهجمات التي بناها المؤلفون أنفسهم؛ فهي تصف مدى نجاح مهاجم قادر يمكنه العمل تحت افتراضات معلنة، لا مدى تكرار الهجمات في العالم الحقيقي. ثانيًا، الأرقام الصارخة — نجاح شبه كامل للهجوم على بعض المرضى — تصف الأفراد الأسوأ حالًا بحكم التصميم؛ وهذا هو الهدف كله، وينبغي قراءتها بوصفها «الذيل أثقل كثيرًا مما يوحي به المتوسط»، لا «معظم المرضى مكشوفون».

الموقف المناسب ليس الذعر ولا التجاهل: دراسة حذرة متعددة البيانات تُظهر أن الطريقة القياسية التي نصادق بها على خصوصية الذكاء الاصطناعي الطبي عمياء تجاه أسوأ حالاتها — وأن هذه الحالات تقع على المرضى الذين لديهم أقل هامش أمان.

لماذا يهم ذلك

هناك سببان يجعلان هذه النتيجة أكثر من هامش تقني.

أولًا، تغيّر ما ينبغي أن يُسمح لعبارة «يحافظ على الخصوصية» أن تعنيه. إذا أُطلق نموذج مع درجة خصوصية متوسطة، يمكن أن تكون هذه الدرجة منخفضة فعلًا وتظل تخفي مجموعة من المرضى يمكن التعرف على عضويتهم بهجوم بصورة شبه كاملة. مطلب الورقة العملي محدد: قيّم خطر الخصوصية **لكل مريض** قبل النشر، واضبط من يستطيع الوصول إلى النماذج المنشورة، واستخدم تقنيات مثل **الخصوصية التفاضلية** — ضجيج صغير مضبوط بعناية يُضاف أثناء التدريب فيخفف هجمات العضوية، مع تكلفة حقيقية ومدارة على فائدة النموذج (مقايضة الخصوصية والمنفعة صريحة وليست مجانية). ينبغي أن تتوقف عبارة «فحصنا المتوسط» عن أن تُحسب كأنها فحص كافٍ.

ثانيًا، تربط الخصوصية بالإنصاف. المجموعات نفسها التي تمثل تمثيلًا ناقصًا في البيانات الطبية — ولذلك تخدمها نماذج الذكاء الاصطناعي الطبي بصورة أسوأ أصلًا — هي التي يتضح أن خصوصيتها محمية أقل. مجال يعمل بجد على التفاوت الأول لا يمكنه التعامل مع الثاني كأنه مشكلة تخص قسمًا آخر. إنهم الأشخاص أنفسهم.

القصة المطمئنة عن خصوصية الذكاء الاصطناعي الطبي — قسناها، المتوسط منخفض، إذن نحن بخير — هي القصة التي تفككها هذه الورقة. ليس لإخافة الناس من التقنية، بل لنقل المعيار إلى مكانه الصحيح: وعد لا تستطيع الادعاء بأنك تفي به إلا إذا فحصته بالنسبة إلى الشخص الأكثر احتمالًا أن يتضرر.

خلاصة واضحة

تحاول هجمات استدلال العضوية تحديد ما إذا كان سجل شخص بعينه ضمن بيانات تدريب نموذج ذكاء اصطناعي — ولأن العضوية نفسها قد تكشف معلومة (مثل أنك كنت مصابًا بمرض معين)، فهي تسرب حقيقي للخصوصية حتى إذا لم يظهر السجل الأصلي أبدًا. كانت الأعمال السابقة تقيس مدى نجاح هذه الهجمات في المتوسط عبر مجموعة بيانات. أما هذه الدراسة فقاستها لكل مريض، عبر سبع مجموعات بيانات طبية (تصوير، ECG، سجلات إلكترونية) ونماذج عديدة لكل منها، ووجدت أن المقاييس الإجمالية تقلل الخطر بشدة: يمكن التعرف على بعض الأفراد تقريبًا بصورة كاملة (AUC ≥ 0.95) للهجوم حتى عندما يبدو المتوسط آمنًا، والأكثر عرضة هم بصورة منهجية من المجموعات ناقصة التمثيل (أقليات عرقية، أمراض نادرة، تصوير غير معتاد)؛ والمشكلة تزداد مع حجم النموذج. لا يجادل المؤلفون بترك الذكاء الاصطناعي الطبي — بل بقياس الخصوصية على مستوى الفرد، وضبط الوصول إلى النموذج، واستخدام الخصوصية التفاضلية. الخلاصة: «خاص في المتوسط» ليس ضمان خصوصية، والأشخاص الذين يفشل معهم هم غالبًا الأقل حماية أصلًا.

مراجعة بلا تهويل

ما الذي تُظهره الورقة: عبر سبع مجموعات بيانات طبية ونماذج كثيرة لكل منها، يكون خطر استدلال العضوية على مستوى المريض أعلى بكثير لبعض الأفراد مما توحي به المقاييس الإجمالية — حتى نجاح شبه مثالي للهجوم (AUC ≥ 95.0) — ويقع هذا الخطر المتبقي بصورة غير متناسبة على المجموعات ناقصة التمثيل ويزداد مع قدرة النموذج.

ما هو معقول لكنه غير مثبت: أن مهاجمين في العالم الحقيقي يستغلون هذا بالفعل ضد نماذج سريرية منشورة؛ الدراسة تثبت القدرة وتوزيعها تحت افتراضات محددة، لا معدل الهجمات في الواقع.

ما الذي لا تُظهره: أن الذكاء الاصطناعي الطبي ينبغي التخلي عنه؛ أو أن كل نموذج يسرّب أو أن نموذجًا منشورًا بعينه اختُرق؛ أو أن محتويات سجل المريض (لا مجرد العضوية) مكشوفة؛ أو رقم تفاوت واحد يلخص كل شيء — النتيجة المتينة هي النمط، لا رقم واحد.

أهم القيود: تقيس خطر أسوأ حالة عبر هجمات بناها المؤلفون، لا حوادث مرصودة؛ والأرقام الدرامية تصف الأكثر انكشافًا بحكم التصميم؛ كما تُعرض الفروق بين المجموعات كنمط متسق عبر البيانات بدل اختزالها في رقم واحد.

ما مقدار الثقة المناسب للقارئ العام؟ ثقة عالية بأن المقاييس الإجمالية للخصوصية تقلل خطر الفرد، وأن الخطر المتبقي يتركز في المرضى ناقصي التمثيل، وأن هذا يسوء مع حجم النموذج — فقد أُظهر عبر بيانات ونماذج كثيرة. وثقة متوسطة بشأن مدى شيوع هذه الهجمات في العالم الحقيقي، وهو ما لا تقيسه الدراسة. القراءة الآمنة ليست «الذكاء الاصطناعي الطبي يسرّب بياناتك» ولا «الخصوصية محلولة»، بل «فحص الخصوصية القياسي أعمى تجاه أسوأ حالاته، وهذه الحالات تقع على أكثر المرضى هشاشة — لذلك يجب أن يتغير الفحص».

المصادر

(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