The Clean Paper

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

Una IA medica puede ser privada en promedio y aun asi exponer a pacientes concretos, sobre todo a los subrepresentados

Un modelo de IA medica llamado preservador de privacidad suele apoyarse en un numero promedio. Este estudio argumenta que ese es el test equivocado. Al estudiar ataques de inferencia de pertenencia - que revelan si el registro de una persona concreta estuvo en los datos de entrenamiento de un modelo y por tanto pueden delatar que tuvo cierta enfermedad -, los autores midieron el riesgo por paciente y no en agregado, en siete conjuntos de datos medicos y muchos modelos. El patron: modelos que parecen seguros en promedio aun pueden permitir identificar a individuos concretos casi perfectamente (AUC de ataque >= 0,95); los expuestos pertenecen sistematicamente a grupos subrepresentados; y empeora a medida que crecen los modelos. No dicen abandonar la IA medica: dicen medir la privacidad por paciente, controlar el acceso al modelo y usar privacidad diferencial.

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

Una garantía de privacidad es una promesa a una persona, no a un promedio

Cuando un modelo de IA médica se llama “preservador de privacidad”, la afirmación suele apoyarse en un número: entre todos los pacientes cuyos datos entrenaron el modelo, la probabilidad promedio de inferir la pertenencia de cualquiera es baja. Suena tranquilizador. El punto discreto e incómodo de este artículo es que ese es el número equivocado.

La privacidad no es un promedio. Es una promesa hecha a cada individuo: que estar en el conjunto de entrenamiento no volvera para exponerlo. Y un promedio puede cumplir esa promesa para casi todos mientras la rompe por completo para unos pocos. Los investigadores midieron el riesgo paciente por paciente, con una resolución que no se había usado antes, y encontraron exactamente eso: modelos que parecen seguros en agregado pueden filtrar casi perfectamente la pertenencia de individuos concretos, y los filtrados son desproporcionadamente quienes ya están menos protegidos.

Qué hicieron los autores

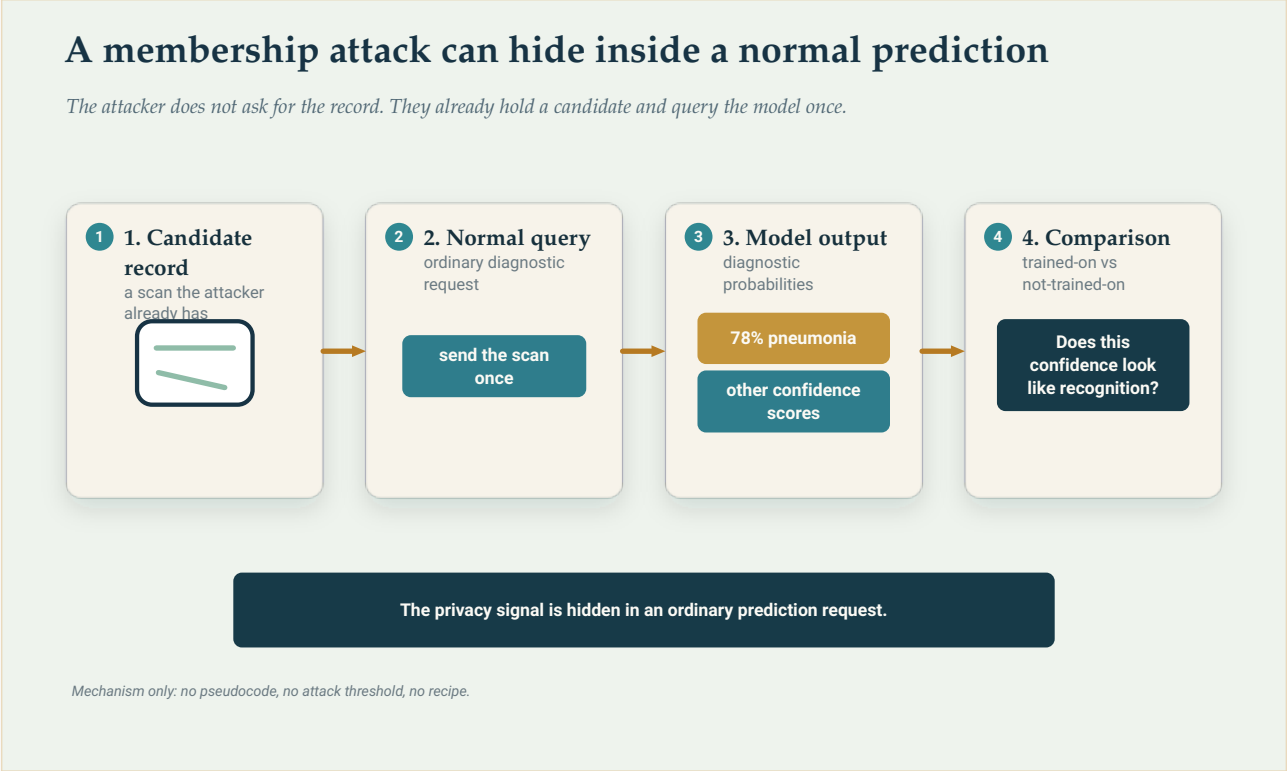
El equipo, liderado por Moritz Knolle, con Daniel Rückert, Georg Kaissis y colegas, tomó una amenaza conocida llamada **ataque de inferencia de pertenencia** y cambió la pregunta. En vez de preguntar “en promedio, con qué frecuencia tiene éxito el ataque en el conjunto?”, preguntaron “para *este paciente concreto*, ¿hasta qué punto está expuesto?”

Ejecutaron ese análisis por paciente en **siete conjuntos de datos médicos establecidos**, de tipos muy distintos: imagen médica, electrocardiogramas e historias clínicas electrónicas, con 200 modelos por conjunto. Para cada paciente en cada modelo, estimaron con qué confianza un atacante podía decir si su registro había formado parte del entrenamiento, y luego desglosaron los resultados por grupo: estado de enfermedad, raza autodeclarada, seguro, sexo y protocolo de imagen.

Qué es realmente un ataque de inferencia de pertenencia

La fuga aquí es más sutil que “el modelo escupe tu registro”. Trata de *pertenencia*: el mero hecho de que tus datos estuvieran en el conjunto de entrenamiento.
Un modelo médico normal recibe una radiografia de torax y devuelve probabilidades: neumonia, edema, consolidacion. Esa respuesta diagnóstica cotidiana es lo único que usa el ataque. Nada exótico.
La debilidad es esta: un modelo suele estar un poco **más confiado** con los ejemplos exactos que vio en entrenamiento que con los que nunca vio. Como un estudiante que vio el examen antes: en las preguntas practicadas responde demasiado rápido y seguro. Deducir, a partir de esa pista, si un registro particular fue uno de los ejemplos estudiados es inferencia de pertenencia.
Así funciona. Alguien quiere saber si el escaneó de una persona se uso para entrenar el modelo. No pide el registro al modelo; ya tiene un escaneó candidato. Lo envía como una consulta diagnóstica ordinaria y mira cuánta confianza tiene la respuesta. Luego compara esa confianza

> con lo que esperaría de un modelo que sí lo hubiera visto o que no, comparación que puede aprender entrenando un modelo de referencia. Si el modelo real está inusualmente seguro sobre ese escaneó, como si lo reconociera, eso es evidencia fuerte de pertenencia.
> Lo inquietante es que la solicitud no parece un ataque: es la misma consulta diagnóstica que haría un clínico, y la señal de privacidad está escondida dentro de una predicción normal. Por eso el riesgo es concreto, aunque hasta ahora se haya demostrado bajo supuestos de laboratorio y no capturado en la naturaleza.
> ¿Por qué importa la pertenencia si el contenido no sale? Porque *la pertenencia es un hecho*. Si un modelo fue entrenado con pacientes que recibieron cierta inmunoterapia contra cáncer, confirmar que tu registro está dentro revela probablemente ese cáncer. El contenido queda sellado; el hecho de pertenecer se escapa.



El ataque no necesita una API especial de privacidad. Una consulta diagnóstica normal puede filtrar una señal de pertenencia si el modelo esta inusualmente confiado con un registro que ya vio.

Qué encontraron

Tres hallazgos, cada uno más agudo que el anterior.

Los promedios esconden a los expuestos. Medidos en agregado, como se hace normalmente, mu-

chos modelos parecen tranquilizadoramente privados: el ataque no supera al azar para la mayoría de pacientes. La vista por paciente contó otra historia. Para algunos individuos, el ataque tiene éxito **casi perfecto**: los autores lo miden como AUC de ataque de **0,95 o más** (0,5 es azar, 1,0 perfecto), incluso cuando el promedio del conjunto parece azar.

El promedio puede quedar cerca del azar mientras una pequeña cola de registros sigue mucho más expuesta. El punto del artículo es que la privacidad debe comprobarse a nivel de paciente, no solo en agregado.

Original Aurora diagram — The Clean Paper CC BY 4.0

La exposición es desigual y cae sobre los subrepresentados. Los pacientes más vulnerables a ataques casi perfectos pertenecen sistemáticamente a grupos **subrepresentados en los datos**: minorias etnicas, fenotipos de enfermedad rara, rasgos de imagen inusuales. En el conjunto de historias electrónicas, pacientes negros aparecieron **31 % más a menudo** de lo esperado entre los registros más vulnerables; en mamografia, escaneos sospechosos de malignidad estaban sobrerrepresentados en esa zona de riesgo por **+1.179 %**. Son ejemplos y sobrerrepresentaciones relativas dentro de la pequeña cola extrema, no la fracción de todos esos pacientes expuestos. El modelo tiene menos ejemplos parecidos para difuminarlos; destacan, y destacar es justo lo que detecta un ataque de pertenencia.

Los modelos más grandes lo empeoran. El número de pacientes expuestos a ataques casi perfectos subió con la capacidad del modelo. En un conjunto de dermatologia, la proporción paso de practicamente cero en el modelo más pequeño a alrededor de **uno de cada diez** en el mayor. A medida que la IA médica se vuelve más grande y capaz, este riesgo específico se agrava.

El resumen de Rückert es seco: “This is not a tolerable risk. Health data is highly sensitive.”

Por qué “seguro en promedio” es el test equivocado

La tentación es leer un riesgo promedio bajo como certificado de seguridad. La lección central es que promediar aquí no es solo impreciso: mide la cosa equivocada.

Una garantía de privacidad solo tiene sentido si se cumple para la persona con más riesgo, no para la persona del medio. Un modelo donde 999 de 1.000 pacientes son irrecuperables pero uno puede identificarse casi perfectamente no es “99,9 % privado” para esa persona; y si esa persona es previsiblemente el paciente de enfermedad rara o de una minoría étnica, la métrica no es solo incompleta, sino discriminatoria. Informa seguridad para la mayoría y la llama seguridad para todos.

El cambio que fuerza el artículo es pasar de *¿hasta qué punto es privado este modelo en promedio?* a *¿quién es el paciente más expuesto y quién es?* Son preguntas distintas, y solo la segunda es una pregunta de privacidad.

Lo que esto no prueba ni afirma

- No dice que haya que abandonar la IA médica. El marco es mitigación, no retirada.
- No significa que todos los modelos médicos filtren, ni que un modelo desplegado concreto haya sido atacado. Muestra que el *riesgo existe y está distribuido de forma desigual*, y que las métricas agregadas lo pierden.
- No significa que tus registros ya estén expuestos. El ataque necesita condiciones específicas: acceso al modelo, un registro candidato e infraestructura de referencia.
- No muestra que filtre el *contenido* del paciente. Filtra pertenencia, peligrosa por otra razón.
- No reduce las disparidades a un solo número. El resultado fuerte es el *patrón*: las métricas agregadas subestiman riesgo individual, y los subrepresentados cargan más.

¿Qué solidez tienen las pruebas?

Es un resultado empírico y metodológico robusto. El patrón - las métricas agregadas subestiman el riesgo por paciente, el riesgo residual se concentra en grupos subrepresentados y empeora con la capacidad - se sostuvo en **siete conjuntos de datos de tipos distintos** y muchos modelos. Esa amplitud lo vuelve una afirmación de medición creible, no un artefacto de un dataset.

Dos salvedades honestas. Primero, es una demostración de *riesgo*, medida ejecutando ataques construidos por los autores; caracteriza lo que un atacante capaz *podría* hacer bajo supuestos, no cuántas veces ocurren ataques reales. Segundo, los números llamativos describen a los individuos peor expuestos *por diseño*; ese es el punto, y deben leerse como “la cola es mucho más pesada de lo que implica el promedio”, no como “la mayoría de pacientes está expuesta”.

La postura adecuada no es alarma ni desden: un estudio cuidadoso y multidataset que muestra que la forma estándar de certificar privacidad en IA médica es ciega a sus peores casos, y que esos casos

caen sobre pacientes con menos margen.

Por qué importa

Dos cosas lo hacen más que una nota técnica.

Primero, cambia lo que debería poder significar "preservador de privacidad". Si un modelo se publica con una puntuación promedio de privacidad, esa puntuación puede ser realmente baja y ocultar aun así un subconjunto de pacientes casi perfectamente identificables. La demanda práctica es concreta: evaluar el riesgo **por paciente** antes de liberar, controlar quién accede a modelos desplegados y usar técnicas como **privacidad diferencial**: ruido pequeño y calibrado durante el entrenamiento que debilita ataques de pertenencia, con un coste real y gestionado para utilidad. "Comprobamos el promedio" debería dejar de contar como comprobación suficiente.

Segundo, une privacidad con equidad. Los mismos grupos subrepresentados en datos médicos, y por tanto peor servidos por IA médica, son aquellos cuya privacidad esa IA protege menos. Un campo que trabaja en una inequidad no puede tratar la otra como problema ajeno. Son las mismas personas.

La historia tranquilizadora - *lo medimos, el promedio es bajo, estamos bien* - es la que este artículo desmonta. No para espantar a nadie de la tecnología, sino para mover el estándar a donde pertenece: una promesa que solo puedes afirmar que cumples si la comprobaste para la persona más probable de ser dañada.

En pocas palabras

Los ataques de inferencia de pertenencia intentan determinar si el registro de una persona concreta estuvo en los datos de entrenamiento de un modelo, y esa pertenencia puede revelar información sensible aunque el registro no salga. Trabajos previos medían el éxito *promedio* de esos ataques. Este estudio lo midió *por paciente*, en siete conjuntos de datos médicos y muchos modelos, y encontró que las métricas agregadas subestiman mucho el riesgo: algunos individuos son identificables casi perfectamente (AUC >= 0,95) aunque el promedio parezca seguro; los más vulnerables vienen de grupos subrepresentados; y el problema crece con el tamaño del modelo. Los autores no piden abandonar la IA médica: piden medir privacidad individual, controlar acceso y usar privacidad diferencial. La conclusión: "privado en promedio" no es una garantía de privacidad.

Sin rodeos

Lo que muestra el artículo: En siete conjuntos de datos médicos y muchos modelos, el riesgo de inferencia de pertenencia por paciente es mucho mayor para algunos individuos que lo que sugieren las métricas agregadas, hasta éxito casi perfecto (AUC >= 0,95), y ese riesgo residual cae despropor-

cionadamente en grupos subrepresentados y crece con la capacidad del modelo.

Lo plausible pero no probado: Qué atacantes reales ya exploten esto contra modelos clínicos desplegados; el estudio demuestra la capacidad y distribución bajo supuestos, no la frecuencia en la práctica.

Lo que no muestra: Qué haya que abandonar la IA médica; que todos los modelos filtren o que uno específico haya sido vulnerado; que se exponga el contenido de registros; una cifra única de disparidad.

Limitaciones principales: Mide riesgo con los ataques de los autores, no incidentes observados; los números dramáticos describen a los peor expuestos por diseño; las disparidades se muestran como patrón consistente, no como un número universal.

¿Cuánta confianza debería tener un lector general? Alta en que las métricas agregadas subestiman riesgo individual, que el riesgo residual se concentra en pacientes subrepresentados y que empeora con el tamaño del modelo. Moderada sobre la frecuencia real de ataques. Lectura segura: no “la IA médica filtra tus datos” ni “la privacidad está resuelta”, sino “el test estándar es ciego a sus peores casos”.

Fuentes

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