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 · 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 garantia de privacidad es una promesa a una persona, no a un promedio

Cuando un modelo de IA medica se llama “preservador de privacidad”, la afirmacion suele apoyarse en un numero: 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 incomodo de este articulo es que ese es el numero 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 resolucion que no se habia 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 estan menos protegidos.

Que hicieron los autores

El equipo, liderado por Moritz Knolle, con Daniel Rückert, Georg Kaissis y colegas, tomo una amenaza conocida llamada **ataque de inferencia de pertenencia** y cambio la pregunta. En vez de preguntar “en promedio, con que frecuencia tiene exito el ataque en el conjunto?”, preguntaron “para *este paciente concreto*, que tan expuesto esta?”

Ejecutaron ese analisis por paciente en **siete conjuntos de datos medicos establecidos**, de tipos muy distintos: imagen medica, electrocardiogramas e historias clinicas electronicas, con 200 modelos por conjunto. Para cada paciente en cada modelo, estimaron con que confianza un atacante podia decir si su registro habia formado parte del entrenamiento, y luego desglosaron los resultados por grupo: estado de enfermedad, raza autodeclarada, seguro, sexo y protocolo de imagen.

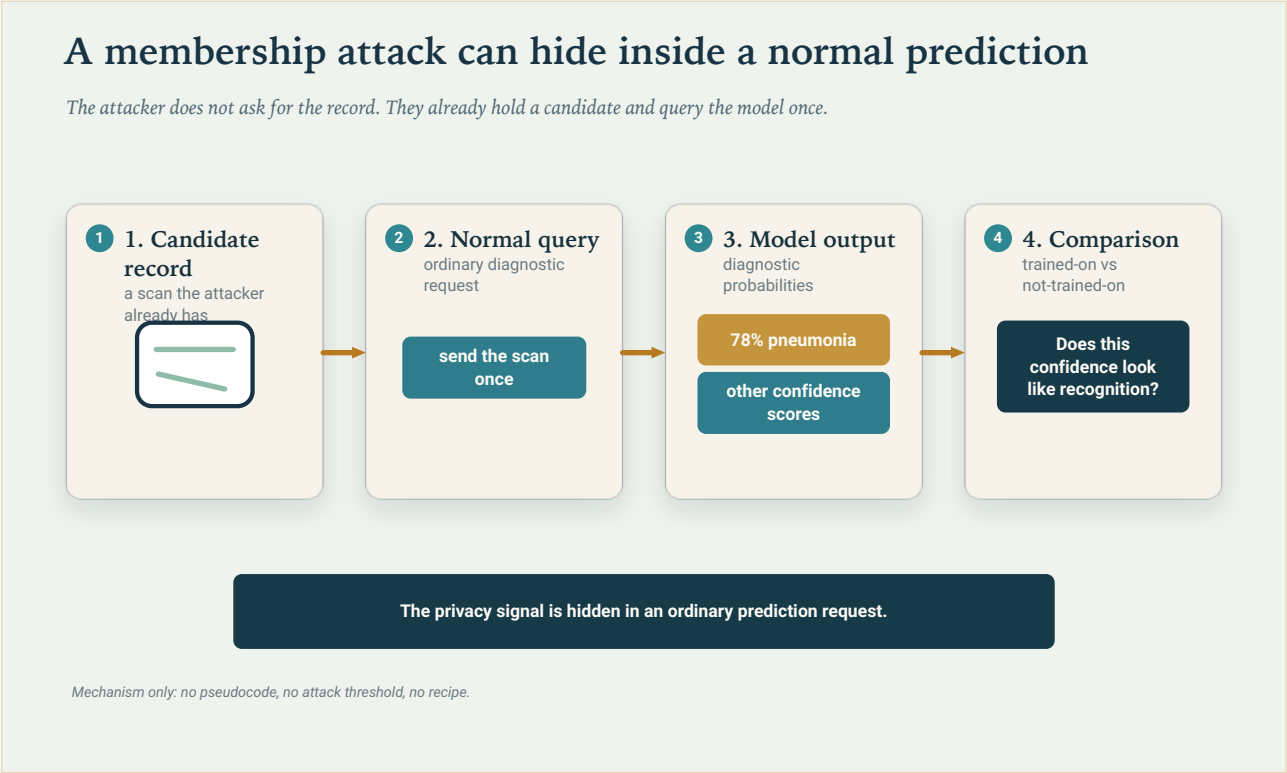
Que es realmente un ataque de inferencia de pertenencia

La fuga aqui es mas 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 medico normal recibe una radiografia de torax y devuelve probabilidades: neumonia, edema, consolidacion. Esa respuesta diagnostica cotidiana es lo unico que usa el ataque. Nada exotico.

La debilidad es esta: un modelo suele estar un poco **mas 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 rapido y seguro. Deducir, a partir de esa pista, si un registro particular fue uno de los ejemplos estudiados es inferencia de pertenencia. Asi funciona. Alguien quiere saber si el escaneo de una persona se uso para entrenar el modelo. No pide el registro al modelo; ya tiene un escaneo candidato. Lo envia como una consulta diagnostica ordinaria y mira cuanta confianza tiene la respuesta. Luego compara esa confianza

> con lo que esperaria de un modelo que si lo hubiera visto o que no, comparacion que puede aprender entrenando un modelo de referencia. Si el modelo real esta inusualmente seguro sobre ese escaneo, como si lo reconociera, eso es evidencia fuerte de pertenencia.
> Lo inquietante es que la solicitud no parece un ataque: es la misma consulta diagnostica que haria un clinico, y la senal de privacidad esta escondida dentro de una prediccion normal. Por eso el riesgo es concreto, aunque hasta ahora se haya demostrado bajo supuestos de laboratorio y no capturado en la naturaleza.
> Por que 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 cancer, confirmar que tu registro esta dentro revela probablemente ese cancer. El contenido queda sellado; el hecho de pertenecer se escapa.



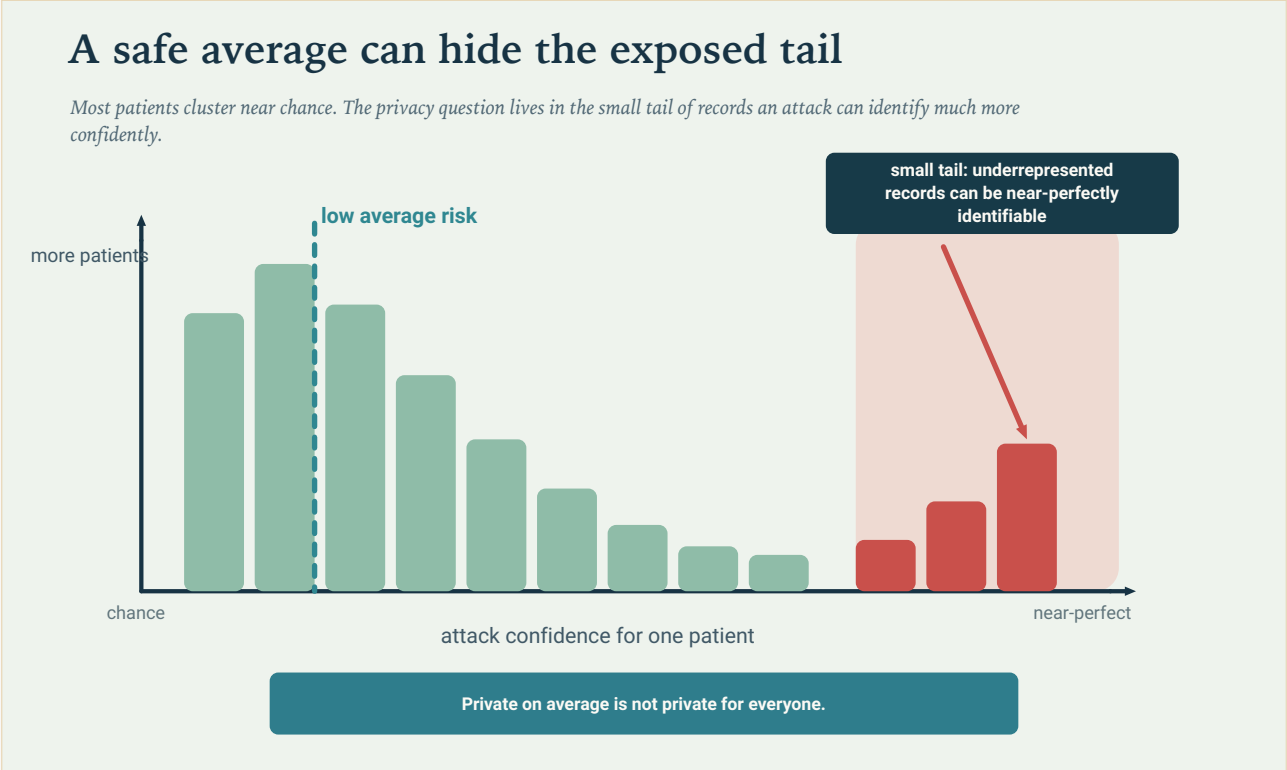
El ataque no necesita una API especial de privacidad. Una consulta diagnostica normal puede filtrar una senal de pertenencia si el modelo esta inusualmente confiado con un registro que ya vio.

Que encontraron

Tres hallazgos, cada uno mas 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 mayoria de pacientes. La vista por paciente conto otra historia. Para algunos individuos, el ataque tiene exito **casi perfecto**: los autores lo miden como AUC de ataque de **0,95 o mas** (0,5 es azar, 1,0 perfecto), incluso cuando el promedio del conjunto parece azar.



El promedio puede quedar cerca del azar mientras una pequena cola de registros sigue mucho mas expuesta. El punto del articulo 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 exposicion es desigual y cae sobre los subrepresentados. Los pacientes mas vulnerables a ataques casi perfectos pertenecen sistematicamente a grupos **subrepresentados en los datos**: minorias etnicas, fenotipos de enfermedad rara, rasgos de imagen inusuales. En el conjunto de historias electronicas, pacientes negros aparecieron **31 % mas a menudo** de lo esperado entre los registros mas 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 pequena cola extrema, no la fraccion 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 mas grandes lo empeoran. El numero de pacientes expuestos a ataques casi perfectos subio con la capacidad del modelo. En un conjunto de dermatologia, la proporcion paso de practicamente cero en el modelo mas pequeno a alrededor de **uno de cada diez** en el mayor. A medida que la IA medica se vuelve mas grande y capaz, este riesgo especifico se agrava.

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

Por que “seguro en promedio” es el test equivocado

La tentacion es leer un riesgo promedio bajo como certificado de seguridad. La leccion central es que promediar aqui no es solo impreciso: mide la cosa equivocada.

Una garantia de privacidad solo tiene sentido si se cumple para la persona con mas 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 minoria etnica, la metrica no es solo incompleta, sino discriminatoria. Informa seguridad para la mayoria y la llama seguridad para todos.

El cambio que fuerza el articulo es pasar de *que tan privado es este modelo en promedio?* a *quien es el paciente mas expuesto, y quien 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 medica. El marco es mitigacion, no retirada.
- No significa que todos los modelos medicos filtren, ni que un modelo desplegado concreto haya sido atacado. Muestra que el *riesgo existe y esta distribuido de forma desigual*, y que las metricas agregadas lo pierden.
- No significa que tus registros ya esten expuestos. El ataque necesita condiciones especificas: acceso al modelo, un registro candidato e infraestructura de referencia.
- No muestra que filtre el *contenido* del paciente. Filtra pertenencia, peligrosa por otra razon.
- No reduce las disparidades a un solo numero. El resultado fuerte es el *patron*: las metricas agregadas subestiman riesgo individual, y los subrepresentados cargan mas.

Que tan fuerte es la evidencia?

Es un resultado empirico y metodologico robusto. El patron - las metricas 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 afirmacion de medicion creible, no un artefacto de un dataset.

Dos salvedades honestas. Primero, es una demostracion de *riesgo*, medida ejecutando ataques construidos por los autores; caracteriza lo que un atacante capaz *podria* hacer bajo supuestos, no cuantas veces ocurren ataques reales. Segundo, los numeros llamativos describen a los individuos peor expuestos *por diseno*; ese es el punto, y deben leerse como “la cola es mucho mas pesada de lo que implica el promedio”, no como “la mayoria de pacientes esta expuesta”.

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

caen sobre pacientes con menos margen.

Por que importa

Dos cosas lo hacen mas que una nota tecnica.

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

Segundo, une privacidad con equidad. Los mismos grupos subrepresentados en datos medicos, y por tanto peor servidos por IA medica, 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 articulo desmonta. No para espantar a nadie de la tecnologia, sino para mover el estandar a donde pertenece: una promesa que solo puedes afirmar que cumples si la comprobaste para la persona mas probable de ser dañada.

Resumen limpio

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 informacion sensible aunque el registro no salga. Trabajos previos median el exito *promedio* de esos ataques. Este estudio lo midio *por paciente*, en siete conjuntos de datos medicos y muchos modelos, y encontro que las metricas agregadas subestiman mucho el riesgo: algunos individuos son identificables casi perfectamente (AUC >= 0,95) aunque el promedio parezca seguro; los mas vulnerables vienen de grupos subrepresentados; y el problema crece con el tamano del modelo. Los autores no piden abandonar la IA medica: piden medir privacidad individual, controlar acceso y usar privacidad diferencial. La conclusion: "privado en promedio" no es una garantia de privacidad.

No-BS check

Lo que muestra el articulo: En siete conjuntos de datos medicos y muchos modelos, el riesgo de inferencia de pertenencia por paciente es mucho mayor para algunos individuos que lo que sugieren las metricas agregadas, hasta exito 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: Que atacantes reales ya exploten esto contra modelos clinicos desplegados; el estudio demuestra la capacidad y distribucion bajo supuestos, no la frecuencia en la practica.

Lo que no muestra: Que haya que abandonar la IA medica; que todos los modelos filtren o que uno especifico haya sido vulnerado; que se exponga el contenido de registros; una cifra unica de disparidad.

Limitaciones principales: Mide riesgo con los ataques de los autores, no incidentes observados; los numeros dramaticos describen a los peor expuestos por diseno; las disparidades se muestran como patron consistente, no como un numero universal.

Cuanta confianza deberia tener un lector general? Alta en que las metricas agregadas subestiman riesgo individual, que el riesgo residual se concentra en pacientes subrepresentados y que empeora con el tamano del modelo. Moderada sobre la frecuencia real de ataques. Lectura segura: no “la IA medica filtra tus datos” ni “la privacidad esta resuelta”, sino “el test estandar es ciego a sus peores casos”.

Sources

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