The Clean Paper

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

Una AI medica può essere privata in media e comunque esporre pazienti specifici — soprattutto i sottorappresentati

Un modello di AI medica definito "privacy-preserving" di solito poggia su un numero medio. Questo studio sostiene che quel numero sia il test sbagliato. Studiando membership inference attack — attacchi che rivelano se il record di una persona specifica era nei dati di addestramento di un modello, e quindi possono tradire che quella persona aveva una certa malattia — gli autori hanno misurato il rischio per paziente invece che in aggregato, su sette dataset medici (imaging, ECG, cartelle elettroniche) e molti modelli per ciascuno. Il pattern: modelli che sembrano sicuri in media possono ancora lasciare identificare individui specifici quasi perfettamente (attack AUC $\geq 0{,}95$); gli esposti sono sistematicamente persone di gruppi sottorappresentati (minoranze etniche, malattie rare, imaging insolito); e il problema peggiora quando i modelli crescono. Gli autori non dicono di abbandonare l'AI medica — dicono di misurare la privacy per paziente, controllare l'accesso ai modelli e usare differential privacy. Il nucleo scomodo: "privato in media" non è una garanzia di privacy, e fallisce proprio i pazienti già meno protetti.

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 garanzia di privacy è una promessa a una persona, non a una media

Quando un modello di AI medica viene chiamato “privacy-preserving”, quella promessa di solito poggia su un solo numero: su tutti i pazienti i cui dati hanno addestrato il modello, la probabilità media che si possa inferire l’appartenenza di un individuo al set di addestramento è bassa. Suona rassicurante. Il punto meno appariscente, ma scomodo, di questo lavoro è che è il numero sbagliato.

La privacy non è una media. È una promessa fatta a ogni individuo — che la sua presenza nel set di addestramento non finirà per esporlo. E una media può mantenere quella promessa per quasi tutti e romperla completamente per pochi. I ricercatori hanno misurato il rischio di privacy un paziente alla volta, a una risoluzione mai usata prima, e hanno trovato esattamente questo: modelli che sembrano sicuri in aggregato possono rivelare l’appartenenza di individui specifici al set di addestramento quasi perfettamente — e gli individui che rivelano sono, in modo sproporzionato, quelli già meno protetti.

Cosa hanno fatto gli autori

Il team — guidato da Moritz Knolle, con Daniel Rückert (entrambi alla Technical University of Munich) e Georg Kaissis (Hasso Plattner Institute), insieme a colleghi dell’Imperial College London — ha preso una minaccia privacy nota, chiamata **attacco di inferenza dell’appartenenza** (*membership inference attack*), e ha cambiato la domanda. Invece di chiedere “in media, quanto spesso questo attacco riesce sul set di dati?”, ha chiesto “per *questo paziente specifico*, quanto è esposto?”

Hanno eseguito quell’analisi per-paziente su **sette set di dati medici consolidati**, che coprono tipi di dati molto diversi — imaging medico, elettrocardiogrammi e cartelle cliniche elettroniche — e, su ciascuno, 200 modelli. Per ogni paziente in ogni modello, hanno stimato con quanta sicurezza un attaccante potesse dire se il record di quella persona faceva parte dei dati di addestramento, poi hanno scomposto i risultati per gruppo: stato di malattia, razza auto-riportata, assicurazione, sesso e protocollo di imaging.

> **Che cos’è davvero un attacco di inferenza dell’appartenenza**
>
> La perdita qui è più sottile di “il modello sputa fuori il tuo record”. Riguarda l’*appartenenza* — il semplice fatto che i tuoi dati fossero nel set di addestramento.
> Parti da ciò che uno di questi modelli fa normalmente. Gli dai una scansione — una radiografia del torace, per esempio — e lui restituisce probabilità: 78% di probabilità di polmonite, insieme a letture per cardiomegalia, edema, consolidamento. Quella risposta diagnostica quotidiana è *l’unica* cosa che l’attacco usa. Niente di esotico.
> La debolezza su cui si appoggia è questa: un modello di solito è un po’ **più sicuro** sugli esempi esatti su cui è stato addestrato che su quelli che non ha mai visto. Pensa a uno studente che ha visto di nascosto l’esame prima: sulle domande che ha già praticato, le risposte arrivano un po’ troppo in fretta, un po’ troppo sicure. Capire, da quell’indizio, se un record particolare era

uno degli esempi studiati dal modello è ciò che si intende per «inferenza dell'appartenenza». Ecco quindi l'attacco, passo per passo. Qualcuno vuole sapere se la scansione di una persona specifica è stata usata per addestrare il modello. **Non** chiede al modello il record — ha già una scansione candidata (quella della persona, o una copia vicina). La invia una volta sola, come normale richiesta diagnostica, e annota quanto è sicura la risposta. Poi controlla se quella sicurezza somiglia di più a un modello che *si era* addestrato su quella scansione o a uno che *non* lo aveva fatto — un confronto che può fare a basso costo addestrando un sostituto proprio (un «modello di riferimento») per imparare com'è quell'eccesso di confidenza rivelatore, su un normale computer senza hardware speciale. Se il modello reale è insolitamente sicuro su questa scansione — come se la riconoscesse — questo è un forte indizio che la scansione fosse nei dati di addestramento.

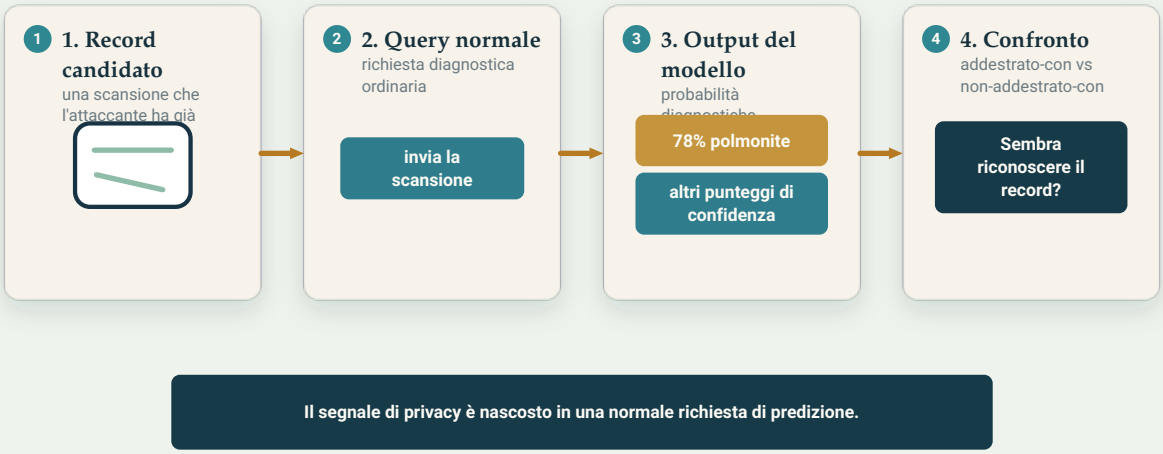
La parte inquietante è che nulla nella richiesta sembra un attacco: è la stessa query diagnostica che farebbe un clinico, e il segnale privacy è nascosto dentro una predizione ordinaria. È proprio per questo che il rischio è concreto — anche se, finora, è stato dimostrato sotto ipotesi di laboratorio dichiarate e non colto in natura.

Perché l'appartenenza conta, se il record stesso non esce mai? Perché *l'appartenenza è un fatto*. Se un modello è stato addestrato su pazienti che hanno ricevuto una certa immunoterapia oncologica, allora confermare che il tuo record è dentro rivela che probabilmente hai avuto quel cancro — il genere di cosa che un assicuratore o un datore di lavoro non dovrebbe mai poter inferire. Il contenuto resta sigillato; è il fatto dell'appartenenza che scappa.

Gli autori espongono queste ipotesi chiaramente — accesso alle predizioni ordinarie del modello, un record candidato e il modello di riferimento dell'attaccante — non come istruzioni operative, ma perché la minaccia possa essere ragionata invece che liquidata.

Un attacco di membership può nascondersi dentro una normale predizione

L'attaccante non chiede il record. Ottiene un candidato e interroga il modello una volta.



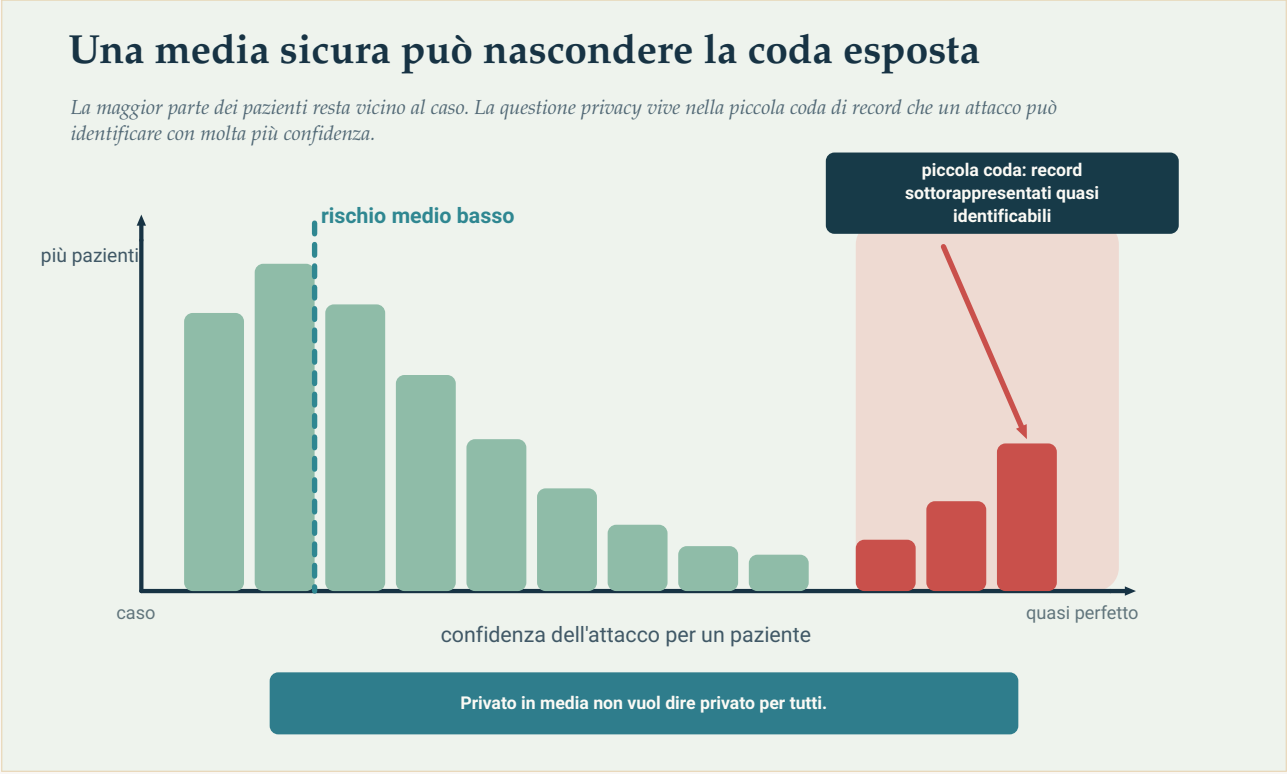
Solo meccanismo: niente pseudocodice, nessuna soglia d'attacco, nessuna ricetta.

L'attacco non ha bisogno di una API privacy speciale. Una normale query diagnostica può rivelare un segnale di appartenenza se il modello è insolitamente sicuro su un record già visto.

Cosa hanno trovato

Tre risultati, ciascuno più netto del precedente.

Le medie nascondono gli esposti. Misurati in aggregato — il modo usuale — molti di questi modelli sembrano rassicuranti: l'attacco non fa meglio del caso per la maggior parte dei pazienti. La vista per-paziente ha raccontato un'altra storia. Come ha detto Knolle nell'annuncio dello studio, le valutazioni precedenti "hanno sempre misurato solo il rischio medio su tutti i pazienti. Noi abbiamo esaminato per la prima volta il rischio a livello di singoli pazienti — e il quadro è molto diverso." Per alcuni individui, l'attacco riesce **quasi perfettamente** — gli autori lo misurano come una attack AUC di **0,95 o superiore** (0,5 è un lancio di moneta, 1,0 è impeccabile) — anche quando la media dell'intero set di dati non sembrava migliore del caso.



La media può stare vicino al caso mentre una piccola coda di record resta molto più esposta. Il punto del lavoro è che la privacy va controllata a livello del paziente, non solo in aggregato.

L'esposizione è diseguale — e colpisce i sottorappresentati. I pazienti più vulnerabili a un attacco quasi perfetto erano sistematicamente quelli in gruppi **sottorappresentati nei dati**: minoranze et-

niche, fenotipi di malattie rare, caratteristiche di imaging insolite. Nel set di dati di cartelle elettroniche, i pazienti neri comparivano **31% più spesso** del previsto tra i record più vulnerabili; nel set di mammografie, le scansioni segnalate come sospette per malignità erano sovrarappresentate in quella zona di pericolo di un impressionante **+1.179%**. (Queste due cifre sono esempi, non l'intero quadro — il lavoro riporta la stessa asimmetria in diversi gruppi — e sono sovrarappresentazioni *relative* dentro la piccola coda di rischio estremo: quanto più spesso questi pazienti compaiono tra i record più esposti, non la frazione di pazienti neri o con mammografia sospetta che sono esposti.) Un modello ha meno esempi simili in cui sfumare questi pazienti, quindi i loro record spiccano — e spiccare è esattamente ciò che un attacco di inferenza dell'appartenenza rileva. Il fallimento privacy non è casuale; si concentra sulle persone già ai margini.

I modelli più grandi peggiorano il problema. Il numero di pazienti esposti ad attacchi quasi perfetti cresceva nettamente con la capacità del modello — in un set di dati dermatologico, la quota saliva da essenzialmente zero nel modello più piccolo a circa **uno su dieci** nel più grande. Mentre i modelli di AI medica diventano più grandi e capaci — la direzione in cui si muove tutto il campo — questo rischio specifico, avvertono gli autori, diventa più grave, non meno.

Il riassunto di Rückert è netto: "This is not a tolerable risk. Health data is highly sensitive."

Perché "sicuro in media" è il test sbagliato

La tentazione è leggere un rischio medio basso come un certificato di buona salute. La lezione centrale del lavoro è che qui fare la media non è soltanto impreciso — misura la cosa sbagliata.

Una garanzia di privacy ha senso solo se regge per la persona più a rischio, non per la persona nel mezzo. Un modello in cui 999 pazienti su 1.000 sono irrecuperabili ma uno può essere identificato quasi perfettamente non è "privato al 99,9%" in alcun senso che conti per quell'unica persona — e se quell'unica persona è prevedibilmente il paziente con malattia rara o il paziente di minoranza etnica, la metrica non è solo incompleta, è silenziosamente discriminatoria. Riporta sicurezza per la maggioranza e la chiama sicurezza per tutti.

Questo è lo spostamento che il lavoro impone: da *quanto è privato questo modello in media?* a *chi è il paziente più esposto, e chi è?* Sono domande diverse, e solo la seconda è una domanda di privacy.

Cosa *non* dimostra — né sostiene

- **Non** dice che l'AI medica vada abbandonata. L'impostazione degli autori è mitigare il rischio, non rinunciare alla tecnologia: misura e correggi il rischio, non smettere di costruire.
- **Non** significa che ogni modello medico stia perdendo dati, o che un qualunque modello in produzione sia stato attaccato. Mostra che il *rischio esiste ed è distribuito in modo diseguale*, e che le metriche aggregate standard lo perdono.
- **Non** significa che i tuoi record siano già esposti. L'attacco richiede condizioni specifiche — accesso

al modello, un record candidato e l'infrastruttura dell'attaccante — non una capacità casuale.
- **Non** mostra che il *contenuto* dei record dei pazienti esca. Ciò che può emergere è l'appartenenza — il fatto dell'inclusione — pericolosa per una ragione diversa, non perché il record venga scaricato.
- **Non** riduce le disparità a un solo numero da titolo. Il risultato forte e riproducibile è lo *schema ricorrente* — le metriche aggregate sottostimano il rischio individuale, e i pazienti sottorappresentati ne portano la quota maggiore — su diversi set di dati e centinaia di modelli.

Quanto sono solide le prove?

È un risultato empirico e metodologico, e robusto. Lo schema ricorrente — le metriche aggregate di privacy sottostimano sistematicamente il rischio per-paziente, il rischio residuo si concentra sui gruppi sottorappresentati e peggiora con la capacità del modello — regge su **sette set di dati di tipi diversi** e su molti modelli per ciascuno. Questa ampiezza è esattamente ciò che rende credibile un'affermazione sulla misurazione, invece che un artefatto di un solo set di dati.

Due cautele importanti. Primo, questa è una dimostrazione di *rischio*, misurata eseguendo gli attacchi costruiti dagli autori stessi; caratterizza quanto bene potrebbe fare un attaccante capace sotto ipotesi dichiarate, non quanto spesso avvengano attacchi reali. Secondo, i numeri vividi — successo quasi perfetto dell'attacco per alcuni pazienti — descrivono per costruzione gli individui messi peggio; questo è il punto, e vanno letti come "la coda è molto più pesante di quanto suggerisca la media", non come "la maggior parte dei pazienti è esposta".

La posizione appropriata non è né allarme né liquidazione: uno studio attento, condotto su più set di dati, che mostra che il modo standard in cui certifichiamo la privacy dell'AI medica è cieco ai propri casi peggiori — e che quei casi peggiori cadono sui pazienti con meno margine da perdere.

Perché conta

Due cose rendono questo più di una nota tecnica.

Primo, cambia ciò che "privacy-preserving" dovrebbe poter significare. Se un modello viene rilasciato con un punteggio medio di privacy, quel punteggio può essere davvero basso e comunque nascondere un sottoinsieme di pazienti quasi perfettamente identificabili da un attacco di inferenza dell'appartenenza. La richiesta pratica del lavoro è concreta: valutare il rischio privacy **per paziente** prima del rilascio, controllare chi può accedere ai modelli distribuiti e usare tecniche come la **privacy differenziale** (*differential privacy*) — piccolo rumore calibrato con cura aggiunto durante l'addestramento, che riduce l'efficacia degli attacchi di inferenza dell'appartenenza a un costo reale e gestito per l'utilità del modello (il compromesso tra privacy e utilità è esplicito, non gratis). "Abbiamo controllato la media" dovrebbe smettere di valere come controllo.

Secondo, intreccia privacy ed equità. Gli stessi gruppi che sono sottorappresentati nei dati medici — e quindi già serviti peggio dall'AI medica — risultano essere quelli la cui privacy quell'AI protegge

meno. Un campo che lavora duramente sulla prima disuguaglianza non può trattare la seconda come affare di qualcun altro. Sono le stesse persone.

La storia rassicurante della privacy nell'AI medica — *l'abbiamo misurata, la media è bassa, siamo a posto* — è la storia che questo lavoro smonta. Non per spaventare qualcuno e allontanarlo dalla tecnologia, ma per spostare lo standard dove deve stare: una promessa che puoi dire di mantenere solo se l'hai controllata per la persona più probabile da ferire.

In breve

Gli attacchi di inferenza dell'appartenenza cercano di determinare se il record di una persona specifica fosse nei dati di addestramento di un modello AI — e poiché l'appartenenza può essere rivelatrice di per sé (che tu abbia avuto una certa malattia, per esempio), è una vera perdita di privacy anche quando il record sottostante non emerge mai. Il lavoro precedente misurava quanto spesso questi attacchi riuscissero *in media* su un set di dati. Questo studio lo ha misurato *per paziente*, su sette set di dati medici (imaging, ECG, cartelle elettroniche) e molti modelli per ciascuno, e ha trovato che le metriche aggregate sottostimano gravemente il rischio: alcuni individui possono essere identificati quasi perfettamente (attack AUC $\geq$ 0,95) anche quando la media dell'intero set di dati sembra sicura; i più vulnerabili sono sistematicamente quelli di gruppi sottorappresentati (minoranze etniche, malattie rare, imaging insolito); e il problema cresce con la dimensione del modello. Gli autori non chiedono di abbandonare l'AI medica: chiedono di misurare la privacy a livello individuale, controllare l'accesso ai modelli e usare la privacy differenziale. Il punto centrale: "privato in media" non è una garanzia di privacy, e le persone che fallisce sono di solito quelle già meno protette.

Valutazione critica

Cosa mostra il lavoro: Su sette set di dati medici e molti modelli per ciascuno, il rischio per-paziente di inferenza dell'appartenenza è molto più alto per alcuni individui di quanto suggeriscano le metriche aggregate — fino a un successo quasi perfetto dell'attacco (AUC $\geq$ 0,95) — e quel rischio residuo cade in modo sproporzionato sui gruppi sottorappresentati e cresce con la capacità del modello.

Cosa è plausibile ma non provato: Che attaccanti reali stiano già sfruttando questo contro modelli clinici distribuiti; lo studio dimostra la *capacità e la sua distribuzione* sotto ipotesi dichiarate, non la frequenza degli attacchi in natura.

Cosa non mostra: Che l'AI medica vada abbandonata; che ogni modello perda dati o che un modello specifico in produzione sia stato violato; che i *contenuti* dei record dei pazienti (invece della semplice appartenenza) siano esposti; una singola cifra di disparità — il risultato robusto è lo schema ricorrente, non un numero.

Limiti principali: Misura il rischio nel caso peggiore tramite gli attacchi degli autori stessi, non incidenti osservati; le cifre drammatiche descrivono per costruzione gli individui più esposti; e le

disparità precise per gruppo sono mostrate come schema coerente nei diversi set di dati, non ridotte a un solo numero.

Quanta fiducia dovrebbe avere un lettore generale? Alta che le metriche aggregate di privacy sottostimino il rischio individuale, che il rischio residuo si concentri sui pazienti sottorappresentati e che peggiori con la dimensione del modello — è dimostrato su molti set di dati e modelli. Moderata sulla frequenza reale di questi attacchi, che questo studio non misura. Lettura sicura: non "l'AI medica perde i tuoi dati", e non "la privacy è risolta", ma "il controllo privacy standard è cieco ai propri casi peggiori, e quei casi cadono sui pazienti più vulnerabili — quindi il controllo deve cambiare."

Fonti

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