The Clean Paper

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

Medicinski AI može biti privatan u prosjeku, a ipak izložiti pojedine pacijente — osobito nedovoljno zastupljene

Za model medicinske umjetne inteligencije koji se naziva „privacy-preserving" obično se navodi jedan prosječni broj. Ova studija tvrdi da je to pogrešan test. Proučavajući napade zaključivanja o članstvu — koji otkrivaju je li zapis određene osobe bio u podacima za treniranje i time mogu odati da je imala neku bolest — autori su mjerili rizik po pacijentu, a ne zbirno, preko sedam medicinskih skupova podataka (snimke, EKG, elektronički kartoni) i mnogo modela po skupu. Obrazac je jasan: modeli koji u prosjeku izgledaju sigurni ipak mogu omogućiti gotovo savršenu identifikaciju određenih pojedinaca (AUC napada ≥ 0.95); izloženi su sustavno pripadnici nedovoljno zastupljenih skupina (manjinske etničke skupine, rijetke bolesti, neuobičajene snimke); a problem se pogoršava kako modeli rastu. Autori ne kažu da treba napustiti medicinski AI — traže mjerenje privatnosti po pacijentu, kontrolu pristupa modelima i diferencijalnu privatnost. Neugodna srž: „privatno u prosjeku" nije jamstvo privatnosti i najčešće zakaže kod pacijenata koji su već najslabije zaštićeni.

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

Jamstvo privatnosti obećanje je osobi, a ne prosjeku

Kada se za model medicinske umjetne inteligencije kaže da „čuva privatnost", ta se tvrdnja obično oslanja na jedan broj: gledano preko svih pacijenata čiji su podaci korišteni za treniranje, prosječna je vjerojatnost da se može zaključiti je li neka osoba bila u skupu za treniranje niska. To zvuči umirujuće. Tiha i neugodna poruka ovog rada jest da je to pogrešan broj.

Privatnost nije prosjek. To je obećanje dano svakoj osobi — da činjenica da je bila u skupu za treniranje neće naknadno dovesti do njezina razotkrivanja. A prosjek može to obećanje održati za gotovo sve i potpuno ga prekršiti za nekolicinu. Istraživači su rizik privatnosti mjerili za svakog pacijenta zasebno, u rezoluciji koja prije nije korištena, i pronašli upravo to: modeli koji zbirno izgledaju sigurni mogu gotovo savršeno otkriti članstvo određenih pojedinaca — a oni koje najviše razotkrivaju nerazmjerno su često ljudi koji su već najslabije zaštićeni.

Što su autori napravili

Tim — predvođen Moritzom Knolleom, s Danielom Rückertom (obojica s Tehničkog sveučilišta u Münchenu) i Georgom Kaissisom (Hasso Plattner Institute), uz kolege s Imperial Collegea London — uzeo je poznatu prijetnju privatnosti zvanu **napad zaključivanja o članstvu** (*membership inference attack*) i promijenio pitanje. Umjesto „koliko često ovaj napad u prosjeku uspijeva na cijelom skupu podataka?", pitali su „koliko je izložen *ovaj konkretni pacijent*?"

Takvu su analizu po pacijentu proveli na **sedam etabliranih medicinskih skupova podataka**, koji obuhvaćaju vrlo različite vrste podataka — medicinske snimke, elektrokardiograme i elektroničke zdravstvene kartone — te na 200 modela po skupu. Za svakog pacijenta u svakom modelu procijenili su koliko pouzdano napadač može utvrditi je li zapis te osobe bio dio podataka za treniranje, a zatim su rezultate razložili po skupinama: status bolesti, samoprijavljena rasa, osiguranje, spol i protokol snimanja.

> **Što zapravo znači napad zaključivanja o članstvu**
>
> Curenje je ovdje suptilnije od „model izbaci vaš karton". Riječ je o *članstvu* — samoj činjenici da su vaši podaci bili u skupu za treniranje.
> Krenimo od onoga što takav model inače radi. Date mu snimku — primjerice rendgensku snimku prsnog koša — a on vraća vjerojatnosti: 78% vjerojatnosti pneumonije, uz procjene za kardiomegaliju, edem i konsolidaciju. Napad koristi *samo* taj obični dijagnostički odgovor. Ništa egzotično.
> Slabost na koju se oslanja jest ovo: model je obično malo **sigurniji** u točne primjere na kojima je treniran nego u one koje nikada nije vidio. Zamislite učenika koji je potajno već vidio ispit — na pitanjima koja je prethodno uvježbao odgovori dolaze malo prebrzo i malo previše samouvjereno. Zaključivanje, na temelju tog signala, je li određeni zapis bio među primjerima koje je model učio upravo je ono što znači „membership inference".

Napad izgleda ovako. Netko želi znati je li snimka određene osobe korištena za treniranje modela. **Ne** traži od modela taj zapis — već ima kandidatsku snimku, bilo original te osobe bilo vrlo sličnu kopiju. Pošalje je jednom kao običan dijagnostički upit i zabilježi koliko je model siguran u odgovor. Zatim provjeri izgleda li ta sigurnost više kao odgovor modela koji *jest* trenirao na toj snimci ili modela koji *nije* — usporedbu može napraviti tako da uvježba vlastiti zamjenski „referentni model" i nauči kako izgleda karakteristična pretjerana sigurnost, na običnom računalu bez posebnog hardvera. Ako je stvarni model neobično siguran za tu snimku — kao da je prepoznaje — to je snažan dokaz da je snimka bila u skupu za treniranje.

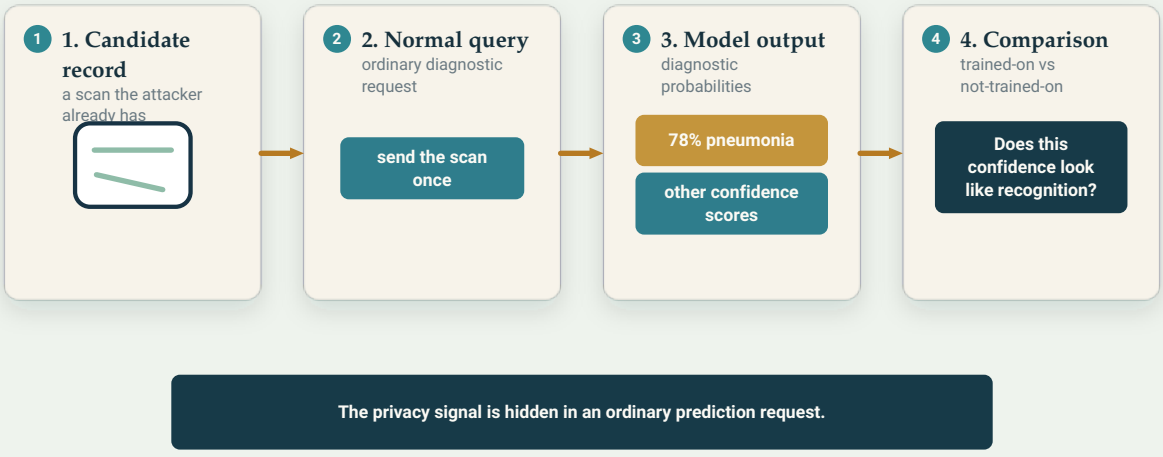
Uznemirujuće je što zahtjev nimalo ne izgleda kao napad: isti je kao običan dijagnostički upit kliničara, a signal privatnosti skriven je u običnom predviđanju. Upravo zato rizik je konkretan — iako je zasad pokazan pod jasno navedenim laboratorijskim pretpostavkama, a ne zabilježen u stvarnom napadu.

Zašto je članstvo važno ako sam zapis nikada ne procuri? Zato što je *članstvo činjenica*. Ako je model treniran na pacijentima koji su primili određenu imunoterapiju za rak, potvrda da je vaš zapis bio u tom skupu može otkriti da ste vjerojatno imali taj rak — upravo onu vrstu podatka koju osiguravatelj ili poslodavac ne bi smio moći zaključiti. Sadržaj zapisa ostaje zatvoren; procuri činjenica pripadnosti.

Autori te pretpostavke otvoreno navode — pristup uobičajenim predviđanjima modela, kandidatski zapis i vlastiti referentni model napadača — ne kao upute za napad, nego kako bi se prijetnja mogla razumno analizirati umjesto površno odbaciti.

A membership attack can hide inside a normal prediction

The attacker does not ask for the record. They already hold a candidate and query the model once.



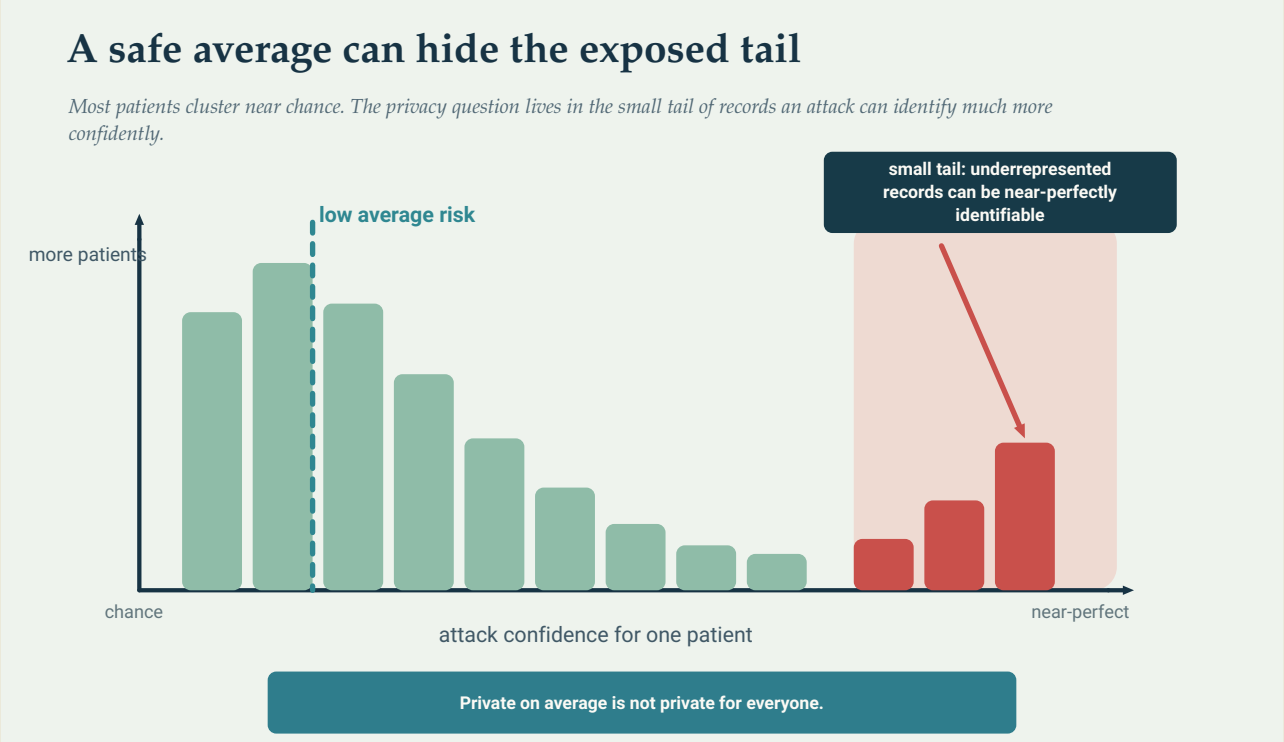
Mechanism only: no pseudocode, no attack threshold, no recipe.

Napadu ne treba poseban API za privatnost. Običan dijagnostički upit može otkriti signal članstva ako je model neobično siguran u zapis koji je već vidio.

Što su pronašli

Tri nalaza, svaki oštriji od prethodnog.

Prosjeci skrivaju izložene pojedince. Kada se mjeri zbirno — kao što je uobičajeno — mnogi od ovih modela izgledaju umirujuće privatno: za većinu pacijenata napad nije bolji od slučajnosti. Pogled na razini pojedinca pokazao je drukčiju sliku. Kako je Knolle rekao u objavi studije, prethodne procjene „uvijek su mjerile samo prosječni rizik preko svih pacijenata. Mi smo prvi put ispitali rizik na razini pojedinačnih pacijenata — i slika je vrlo drukčija." Za neke pojedince napad uspijeva **gotovo savršeno** — autori to mjere kao AUC napada od **0.95 ili više** (0.5 je slučajno pogađanje, 1.0 savršeno) — čak i kada prosjek cijelog skupa izgleda kao slučajnost.



Prosjek može ostati blizu slučajnosti dok mali rep zapisa ostaje mnogo izloženiji. Poanta rada je da privatnost treba provjeravati na razini pacijenta, a ne samo zbirno.

Izloženost nije jednaka — i pogađa nedovoljno zastupljene. Pacijenti najranjiviji na gotovo savršen napad sustavno su bili oni iz skupina **nedovoljno zastupljenih u podacima**: manjinske etničke skupine, rijetki fenotipovi bolesti, neuobičajene značajke snimanja. U skupu elektroničkih kartona crni pacijenti pojavljivali su se **31% češće** nego što bi se očekivalo među najranjivijim zapisima; u skupu mamografija snimke označene kao sumnjive na malignitet bile su u toj opasnoj zoni prekomjerno zastupljene za čak **+1,179%**. (Ta su dva broja primjeri, a ne cijela slika — rad pokazuje

isti pomak u više skupina — i riječ je o *relativnoj* prekomjernoj zastupljenosti unutar malenog repa ekstremnog rizika: koliko se češće ti pacijenti pojavljuju među najizloženijim zapisima, a ne koliki je udio svih crnih pacijenata ili svih sumnjivih mamografija izložen.) Model ima manje sličnih primjera u koje bi takve pacijente mogao „utopiti", pa njihovi zapisi odskaču — a upravo to napad članstva otkriva. Neuspjeh privatnosti nije slučajan; koncentrira se na ljude koji su već na marginama.

Veći modeli pogoršavaju problem. Broj pacijenata izloženih gotovo savršenim napadima naglo je rastao s kapacitetom modela — u jednom dermatološkom skupu udio je porastao od praktički nule kod najmanjeg modela do otprilike **jednog od deset** kod najvećeg. Kako medicinski AI modeli postaju sve veći i sposobniji — a cijelo polje ide u tom smjeru — ovaj specifični rizik, upozoravaju autori, postaje ozbiljniji, ne manji.

Rückertov sažetak je izravan: „To nije prihvatljiv rizik. Zdravstveni podaci izrazito su osjetljivi."

Zašto je „sigurno u prosjeku" pogrešan test

Primamljivo je nizak prosječni rizik pročitati kao potvrdu da je sve u redu. Središnja pouka rada jest da prosjek ovdje nije samo neprecizan — mjeri pogrešnu stvar.

Jamstvo privatnosti ima smisla samo ako vrijedi za osobu s najvećim rizikom, a ne za prosječnu osobu. Model u kojem je 999 od 1,000 pacijenata praktički nemoguće identificirati, ali se jednoga može identificirati gotovo savršeno, nije „99.9% privatan" ni u kojem smislu koji je važan toj jednoj osobi — a ako je ta osoba predvidljivo pacijent s rijetkom bolešću ili pripadnik etničke manjine, metrika nije samo nepotpuna nego i tiho diskriminatorna. Prijavljuje sigurnost većine i naziva je sigurnošću svih.

To je pomak na koji rad prisiljava: od pitanja *koliko je ovaj model u prosjeku privatan?* prema *koji je pacijent najizloženiji i tko je ta osoba?* To nisu ista pitanja, a samo drugo je pravo pitanje privatnosti.

Što ovo *ne* dokazuje — niti tvrdi

- **Ne** kaže da medicinsku umjetnu inteligenciju treba napustiti. Autori govore o ublažavanju rizika, ne povlačenju: izmjeriti i popraviti rizik, a ne prestati graditi.
- **Ne** znači da svaki medicinski model curi ili da je neki konkretni model u uporabi napadnut. Pokazuje da *rizik postoji i da je nejednako raspoređen* te da ga standardne zbirne metrike propuštaju.
- **Ne** znači da su vaši zdravstveni zapisi već izloženi. Napad zahtijeva određene uvjete — pristup modelu, kandidatski zapis i vlastitu infrastrukturu napadača — nije usputna sposobnost.
- **Ne** pokazuje curenje *sadržaja* kartona. Curi članstvo — činjenica uključenosti — što je opasno iz drugog razloga, a ne zato što model izbacuje sam zapis.
- **Ne** svodi nejednakosti na jedan upečatljiv broj. Snažan i ponovljiv nalaz jest *obrazac* — zbirne metrike podcjenjuju individualni rizik, a najveći teret nose nedovoljno zastupljeni pacijenti — preko više skupova podataka i stotina modela.

Koliko su dokazi jaki?

Ovo je empirijski metodološki rezultat i prilično je robustan. Obrazac — zbirne metrike privatnosti sustavno podcjenjuju rizik pojedinca, preostali rizik koncentrira se na nedovoljno zastupljene skupine i pogoršava s kapacitetom modela — održao se na **sedam skupova podataka različitih vrsta** i velikom broju modela po skupu. Upravo ta širina čini mjernu tvrdnju vjerodostojnom umjesto artefaktom jednog skupa.

Dvije poštene ograde. Prvo, ovo je demonstracija *rizika*, mjerena pokretanjem napada koje su sami autori izgradili; opisuje koliko bi sposoban napadač *mogao* uspjeti pod navedenim pretpostavkama, a ne koliko se često napadi događaju u stvarnom svijetu. Drugo, upečatljivi brojevi — gotovo savršen uspjeh napada za neke pacijente — po dizajnu opisuju *najlošije zaštićene pojedince*; to i jest poanta i treba ih čitati kao „rep raspodjele mnogo je teži nego što prosjek sugerira", a ne kao „većina pacijenata je izložena".

Primjeren stav nije ni panika ni odbacivanje: riječ je o pažljivoj studiji na više skupova koja pokazuje da je standardni način potvrđivanja privatnosti medicinskog AI-ja slijep za vlastite najgore slučajeve — i da ti slučajevi padaju upravo na pacijente s najmanje prostora za dodatnu štetu.

Zašto je to važno

Dvije stvari čine ovo više od tehničke fusnote.

Prvo, mijenja što bi izraz „čuva privatnost" smio značiti. Ako se model objavi s prosječnim rezultatom privatnosti, taj broj može biti doista nizak i ipak skrivati podskup pacijenata koje napad članstva može gotovo savršeno identificirati. Praktični zahtjev rada je konkretan: procjenjivati rizik privatnosti **po pacijentu** prije objave, kontrolirati tko može pristupiti modelima u uporabi i primjenjivati tehnike poput **diferencijalne privatnosti** — male, pažljivo kalibrirane količine šuma dodane tijekom treniranja koje otupljuju napade članstva, uz stvaran i upravljiv trošak za korisnost modela (kompromis privatnost–korisnost izričit je, nije besplatan). „Provjerili smo prosjek" više ne bi smjelo značiti da smo provjerili privatnost.

Drugo, povezuje privatnost s pravednošću. Iste skupine koje su nedovoljno zastupljene u medicinskim podacima — i kojima medicinska umjetna inteligencija zbog toga već lošije služi — ispadaju skupine čiju privatnost taj AI najmanje štiti. Polje koje ozbiljno radi na prvoj nejednakosti ne može drugu tretirati kao tuđi problem. Riječ je o istim ljudima.

Umirujuća priča o privatnosti medicinskog AI-ja — *izmjerili smo, prosjek je nizak, sve je u redu* — upravo je priča koju ovaj rad rastavlja. Ne da bi ljude odvratio od tehnologije, nego da bi standard pomaknuo tamo gdje pripada: obećanje za koje smijete tvrditi da ga držite tek nakon što ste ga provjerili za osobu koja će najvjerojatnije biti povrijeđena.

Sažetak bez ukrasa

Napadi zaključivanja o članstvu pokušavaju utvrditi je li zapis određene osobe bio u podacima za treniranje AI modela — a budući da samo članstvo može otkrivati osjetljivu informaciju, primjerice da ste imali određenu bolest, to je stvarno curenje privatnosti čak i kada sam zapis nikada ne izađe iz modela. Prethodni rad mjerio je koliko takvi napadi uspijevaju *u prosjeku* preko skupa podataka. Ova studija mjerila je rizik *po pacijentu*, preko sedam medicinskih skupova podataka (snimke, EKG, elektronički kartoni) i mnogo modela po skupu, te našla da zbirne metrike ozbiljno podcjenjuju rizik: neki pojedinci mogu se identificirati gotovo savršeno (AUC napada ≥ 0.95) čak i kada prosjek cijelog skupa izgleda sigurno; najranjiviji su sustavno pripadnici nedovoljno zastupljenih skupina (manjinska etnička pripadnost, rijetka bolest, neuobičajene snimke); a problem raste s veličinom modela. Autori ne pozivaju na napuštanje medicinskog AI-ja — traže mjerenje privatnosti po pacijentu, kontrolu pristupa modelima i diferencijalnu privatnost. Zaključak: „privatno u prosjeku" nije jamstvo privatnosti, a najčešće zakaže upravo kod pacijenata koji su već najslabije zaštićeni.

Provjera bez uljepšavanja

Što rad pokazuje: Preko sedam medicinskih skupova podataka i mnogih modela po skupu, rizik napada zaključivanja o članstvu po pacijentu za neke je pojedince mnogo veći nego što sugeriraju zbirne metrike — sve do gotovo savršenog uspjeha napada (AUC ≥ 0.95) — a taj preostali rizik nerazmjerno pogađa nedovoljno zastupljene skupine i raste s kapacitetom modela.

Što je vjerojatno, ali nije dokazano: Da napadači u stvarnom svijetu to već iskorištavaju protiv kliničkih modela u uporabi; studija pokazuje *sposobnost i njezinu raspodjelu* pod navedenim pretpostavkama, a ne učestalost napada u praksi.

Što ne pokazuje: Da medicinski AI treba napustiti; da svaki model curi ili da je neki konkretni model već probijen; da su izloženi *sadržaji* kartona, a ne članstvo; jedan jedini broj koji opisuje nejednakost — robustan rezultat je obrazac, a ne jedna brojka.

Glavna ograničenja: Mjeri najgori rizik napadima koje su izgradili sami autori, a ne opaženim incidentima; dramatične vrijednosti po dizajnu opisuju najizloženije pojedince; točne razlike među skupinama prikazane su kao dosljedan obrazac preko skupova podataka, a ne svedene na jednu brojku.

Koliko povjerenja treba imati opći čitatelj? Visoko da zbirne metrike privatnosti podcjenjuju individualni rizik, da se preostali rizik koncentrira na nedovoljno zastupljene pacijente i da se pogoršava s veličinom modela — pokazano je na mnogim skupovima i modelima. Umjereno za učestalost takvih napada u stvarnom svijetu, koju ova studija ne mjeri. Sigurno čitanje nije „medicinski AI vam otkriva podatke", niti „privatnost je riješena", nego: „standardna provjera privatnosti slijepa je za vlastite najgore slučajeve, a oni padaju na najranjivije pacijente — zato se provjera mora promijeniti."

Izvori

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