The Clean Paper

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

Medicinski AI može biti privatan u proseku, a ipak izložiti pojedine pacijente — naročito nedovoljno zastupljene

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

Garancija privatnosti obećanje je osobi, a ne proseku

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

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

Šta su autori uradili

Tim — predvođen Moritzom Knolleom, s Danielom Rückertom (obojica s Tehničkog univerziteta u Münchenu) i Georgom Kaissisom (Hasso Plattner Institute), uz kolege s Imperial Collegea London — uzeo je poznatu pretnju privatnosti zvanu **napad zaključivanja o članstvu** (*membership inference attack*) i promenio pitanje. Umesto „koliko često ovaj napad u proseku uspeva na celom skupu podataka?", pitali su „koliko je izložen *ovaj konkretni pacijent*?"

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

Šta zapravo znači napad zaključivanja o članstvu

Curenje je ovde suptilnije od „model izbaci vaš karton". Reč je o *članstvu* — samoj činjenici da su vaši podaci bili u skupu za treniranje.

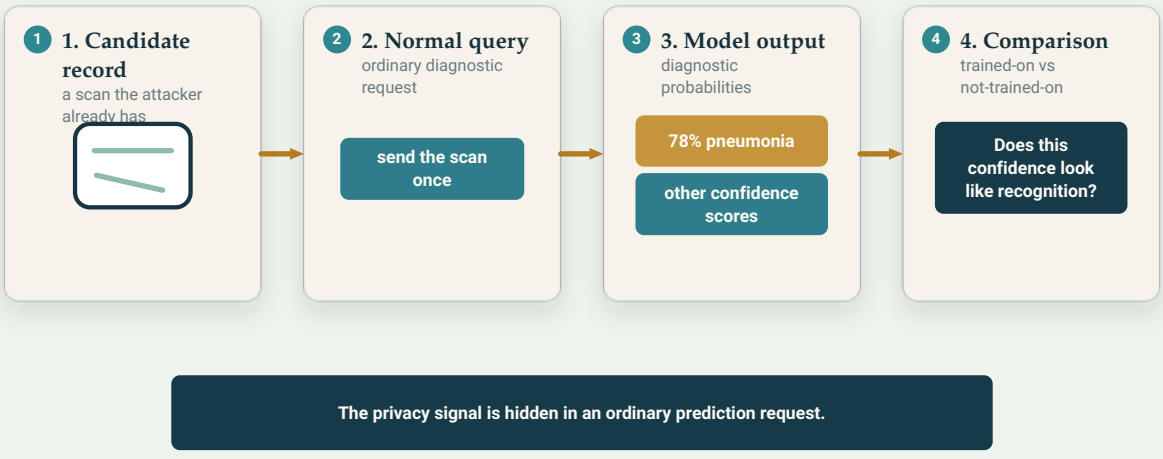
Krenimo od onoga što takav model inače radi. Date mu snimku — primerice rendgensku snimku prsnog koša — a on vraća verovatnoće: 78% verovatnoće pneumonije, uz procene za kardiomegaliju, edem i konsolidaciju. Napad koristi *samo* taj obični dijagnostički odgovor. Ništa egzotično.

Slabost na koju se oslanja je ovo: model je obično malo **sigurniji** u tačne primere 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 samouvereno. Zaključivanje, na osnovu tog signala, da li je određeni zapis bio među primerima koje je model učio upravo je ono što znači „membership inference".

Napad izgleda ovako. Neko želi znati da li je snimka određene osobe korišćena 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 zabeleži koliko je model siguran u odgovor. Zatim proveri izgleda li ta bezbednost više kao odgovor modela koji *je* trenirao na toj snimci ili modela koji *nije* — uporedbu može napraviti tako da uvježba sopstveni zamenski „referentni model" i nauči kako izgleda karakteristična preterana bezbednost, na običnom računaru 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 zahtev 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 zabeležen 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 verovatno imali taj rak — upravo onu vrstu podatka koju osiguravatelj ili poslodavac ne bi smeo da zaključiti. Sadržaj zapisa ostaje zatvoren; procuri činjenica pripadnosti.
Autori te pretpostavke otvoreno navode — pristup uobičajenim predviđanjima modela, kandidatski zapis i sopstveni referentni model napadača — ne kao uputstva za napad, nego kako bi se pretnja mogla razumno analizirati umesto 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.

Šta su pronašli

Tri nalaza, svaki oštriji od prethodnog.

Prosjeci skrivaju izložene pojedince. Kada se meri zbirno — kao što je uobičajeno — mnogi od ovih modela izgledaju umirujuće privatno: za većinu pacijenata napad nije bolji od slučajnosti. Pogled na nivou pojedinca pokazao je drukčiju sliku. Kako je Knolle rekao u objavi studije, prethodne procene „uvek su merile samo prosečni rizik preko svih pacijenata. Mi smo prvi put ispitali rizik na nivou pojedinačnih pacijenata — i slika je vrlo drugačija." Za neke pojedince napad uspeva **gotovo savršeno** — autori to mere kao AUC napada od **0.95 ili više** (0.5 je slučajno pogađanje, 1.0 savršeno) — čak i kada prosek celog skupa izgleda kao slučajnost.

Prosek može ostati blizu slučajnosti dok mali rep zapisa ostaje mnogo izloženiji. Poenta rada je da privatnost treba proveravati na nivou pacijenta, a ne samo zbirno.

Izloženost nije jednaka — i pogađa nedovoljno zastupljene. Pacijenti najranjiviji na gotovo savršen napad sistemno su bili oni iz grupa **nedovoljno zastupljenih u podacima**: manjinske etničke grupe, retki fenotipovi bolesti, neuobičajene funkcije 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 prekodnosno zastupljene za čak **+1,179%**. (Ta su dva broja primeri, a ne cela slika — rad pokazuje isti pomak u

više grupa — i reč je o *relativnoj* prekodnosnoj zastupljenosti unutar malenog repa ekstremnog rizika: koliko se češće ti pacijenti pojavljuju među najizloženijim zapisima, a ne koliki je udeo svih crnih pacijenata ili svih sumnjivih mamografija izložen.) Model ima manje sličnih primera u koje bi takve pacijente mogao „utopiti", pa njihovi zapisi odskaču — a upravo to napad članstva otkriva. Neuspeh 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 udeo je porastao od praktično nule kod najmanjeg modela do otprilike **jednog od deset** kod najvećeg. Kako medicinski AI modeli postaju sve veći i sposobniji — a celo polje ide u tom smeru — ovaj specifični rizik, upozoravaju autori, postaje ozbiljniji, ne manji.

Rückertov rezime je direktan: „To nije prihvatljiv rizik. Zdravstveni podaci izrazito su osetljivi."

Zašto je „sigurno u proseku" pogrešan test

Primamljivo je nizak prosečni rizik pročitati kao potvrdu da je sve u redu. Centralna pouka rada je da prosek ovde nije samo neprecizan — meri pogrešnu stvar.

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

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

Šta ovo *ne* dokazuje — niti tvrdi

- **Ne** kaže da medicinsku veštačku inteligenciju treba napustiti. Autori govore o ublažavanju rizika, ne povlačenju: izmeriti i popraviti rizik, a ne prestati graditi.
- **Ne** znači da svaki medicinski model curi ili da je neki konkretni model u upotrebi 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 zahteva određene uslove — pristup modelu, kandidatski zapis i sopstvenu 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 je *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 sistemno podcjenjuju rizik pojedinca, preostali rizik koncentrira se na nedovoljno zastupljene grupe i pogoršava s kapacitetom modela — održao se na **sedam skupova podataka različitih vrsta** i velikom broju modela po skupu. Upravo ta širina čini mernu tvrdnju verodostojnom umesto artefaktom jednog skupa.

Dve poštene ograde. Prvo, ovo je demonstracija *rizika*, merena pokretanjem napada koje su sami autori izgradili; opisuje koliko bi sposoban napadač *mogao* uspeti pod navedenim pretpostavkama, a ne koliko se često napadi događaju u stvarnom svetu. Drugo, upečatljivi brojevi — gotovo savršen uspeh napada za neke pacijente — po dizajnu opisuju *najlošije zaštićene pojedince*; to i je poenta i treba ih čitati kao „rep raspodele mnogo je teži nego što prosek sugeriše", a ne kao „većina pacijenata je izložena".

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

Zašto je to važno

Dve stvari čine ovo više od tehničke fusnote.

Prvo, menja što bi izraz „čuva privatnost" smio značiti. Ako se model objavi s prosečnim rezultatom privatnosti, taj broj može biti zaista nizak i ipak skrivati podskup pacijenata koje napad članstva može gotovo savršeno identifikovati. Praktični zahtev rada je konkretan: procenjivati rizik privatnosti **po pacijentu** pre objave, kontrolirati ko može pristupiti modelima u upotrebi i primenjivati tehnike poput **diferencijalne privatnosti** — male, pažljivo kalibrirane količine šuma dodane tokom treniranja koje otupljuju napade članstva, uz stvaran i upravljiv trošak za korisnost modela (kompromis privatnost–korisnost izričit je, nije besplatan). „proverili smo prosek" više ne bi smelo značiti da smo proverili privatnost.

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

Umirujuća priča o privatnosti medicinskog AI-ja — *izmerili smo, prosek 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 gde pripada: obećanje za koje smete tvrditi da ga držite tek nakon što ste ga proverili za osobu koja će najverovatnije biti povređena.

Kratak rezime

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

Provera bez ulepšavanja

Šta 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 sugerišu zbirne metrike — sve do gotovo savršenog uspeha napada (AUC ≥ 0.95) — a taj preostali rizik nesrazmerno pogađa nedovoljno zastupljene grupe i raste s kapacitetom modela.

Šta je verovatno, ali nije dokazano: Da napadači u stvarnom svetu to već iskorišćavaju protiv kliničkih modela u upotrebi; studija pokazuje *sposobnost i njenu raspodelu* pod navedenim pretpostavkama, a ne učestalost napada u praksi.

Šta 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: Meri najgori rizik napadima koje su izgradili sami autori, a ne uočenim incidentima; dramatične vrednosti po dizajnu opisuju najizloženije pojedince; tačne razlike među grupama prikazane su kao dosledan obrazac preko skupova podataka, a ne svedene na jednu brojku.

Koliko poverenja treba imati opšti čitalac? 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. Umereno za učestalost takvih napada u stvarnom svetu, koju ova studija ne meri. Sigurno čitanje nije „medicinski AI vam otkriva podatke", niti „privatnost je rešena", nego: „standardna provera privatnosti slepa je za sopstvene najgore slučajeve, a oni padaju na najranjivije pacijente — zato se provera mora promeniti."

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