The Clean Paper

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

Medicinska AI je lahko v povprečju zasebna in hkrati razkrije posamezne bolnike — najbolj prav podzastopane

Trditev, da medicinski model AI »varuje zasebnost«, navadno temelji na eni povprečni številki. Ta študija trdi, da je to napačen preizkus. Avtorji so pri napadih sklepanja o članstvu — ki razkrijejo, ali je bil zapis določene osebe v učnih podatkih modela in s tem lahko izdajo, da je imela določeno bolezen — merili tveganje pri posameznem bolniku, ne le agregatno, na sedmih medicinskih podatkovnih zbirkah (slike, EKG, elektronski zapisi) in številnih modelih. Vzorec: modeli, ki so v povprečju videti varni, lahko še vedno omogočajo skoraj popolno identifikacijo določenih posameznikov (AUC napada ≥ 0,95); izpostavljeni so sistematično ljudje iz podzastopanih skupin (manjšinske etnične skupine, redke bolezni, nenavadne slikovne značilnosti); tveganje pa se povečuje z velikostjo modela. Avtorji ne pravijo, naj medicinsko AI opustimo — zahtevajo merjenje zasebnosti po bolniku, nadzor dostopa do modelov in diferencialno zasebnost. Neprijetno jedro: »zasebno v povprečju« ni jamstvo zasebnosti in odpove prav pri bolnikih, ki so že najmanj zaščiteni.

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 zasebnosti je obljuba posamezniku, ne povprečju

Ko medicinski model AI označimo kot »tak, ki varuje zasebnost«, trditev navadno temelji na eni številki: če pogledamo vse bolnike, katerih podatki so bili uporabljeni za učenje, je povprečna možnost, da kdo ugotovi članstvo posameznega zapisa, majhna. To zveni pomirjujoče. Tiha in neprijetna poanta tega članka je, da je to napačna številka.

Zasebnost ni povprečje. Je obljuba vsakemu posamezniku, da se dejstvo, da je bil v učnem naboru, ne bo vrnilo kot razkritje. Povprečje lahko obljubo drži skoraj vsem, nekaj ljudem pa jo povsem prekrši. Raziskovalci so tveganje zasebnosti izmerili bolnik za bolnikom, z ločljivostjo, ki je prej niso uporabljali, in našli prav to: modeli, ki so agregatno videti varni, lahko članstvo nekaterih posameznikov razkrijejo skoraj popolnoma — in nesorazmerno pogosto razkrijejo prav ljudi, ki so že najmanj zaščiteni.

Kaj so naredili avtorji

Ekipa pod vodstvom Moritza Knolleja, z Danielom Rückertom (oba Tehniška univerza München), Georgom Kaissisom (Hasso Plattner Institute) ter sodelavci z Imperial College London, je vzela znano grožnjo zasebnosti, imenovano **napad sklepanja o članstvu**, in spremenila vprašanje. Namesto »kako pogosto ta napad v povprečju uspe v celotnem podatkovnem naboru?« so vprašali: »kako izpostavljen je *ta konkretni bolnik*?«

Analizo na ravni posameznika so izvedli na **sedmih uveljavljenih medicinskih podatkovnih zbirkah**, ki zajemajo različne tipe podatkov — medicinsko slikanje, elektrokardiograme in elektronske zdravstvene zapise — ter pri vsaki preizkusili 200 modelov. Za vsakega bolnika v vsakem modelu so ocenili, kako zanesljivo lahko napadalec ugotovi, ali je bil bolnikov zapis del učnih podatkov, nato pa rezultate razčlenili po skupinah: bolezni, samoopredeljeni rasi, zavarovanju, spolu in slikovnem protokolu.

Kaj napad sklepanja o članstvu dejansko pomeni

Tukajšnje uhajanje je bolj prefinjeno kot »model izpljune vaš zapis«. Gre za *članstvo* — že samo dejstvo, da so bili vaši podatki v učnem naboru.

Začnimo z običajnim delovanjem modela. Vnesemo sliko, denimo rentgen prsnega koša, model pa vrne verjetnosti: 78 % za pljučnico in druge ocene za kardiomegalijo, edem, konsolidacijo. Napad uporablja samo ta vsakdanji diagnostični odgovor. Nič posebnega.

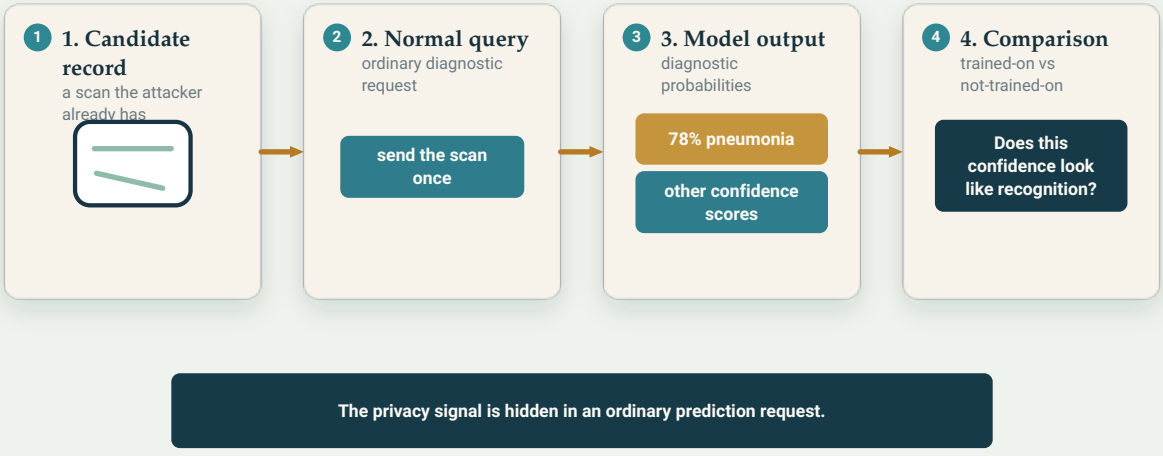
Slabost, ki jo izkorišča, je naslednja: model je običajno nekoliko **bolj samozavesten** pri točno tistih primerih, na katerih se je učil, kot pri primerih, ki jih še ni videl. Kot študent, ki je skrivaj vnaprej videl izpit — pri vajenih vprašanjih odgovor pride malce prehitro in preveč samozavestno. Ugotavljanje, ali je bil določen zapis eden od primerov, ki jih je model »vadil«, na podlagi tega signala imenujemo sklepanje o članstvu.

Napad gre po korakih takole. Nekdo želi izvedeti, ali je bila slika določene osebe uporabljena za

učenje modela. Modela **ne** prosi za zapis — kandidatno sliko že ima (lastno sliko te osebe ali zelo podobno kopijo). Pošlje jo kot navadno diagnostično zahtevo in zabeleži, kako samozavesten je odgovor. Nato preveri, ali ta stopnja samozavesti bolj spominja na model, ki je sliko *videl med učenjem*, ali na model, ki je *ni*. Primerjavo lahko poceni nauči z lastnim nadomestnim oziroma »referenčnim« modelom na običajnem računalniku. Če je pravi model pri določeni sliki nenavadno gotov — kot da jo prepozna — je to močan znak, da je bila v učnem naboru.
Neprijetno je, da zahteva sploh ni videti kot napad: enaka je običajni diagnostični poizvedbi, signal zasebnosti pa je skrit v navadni napovedi. Prav zato je tveganje konkretno, čeprav je bilo doslej pokazano pod navedenimi laboratorijskimi predpostavkami in ne opaženo v divjini.
Zakaj članstvo šteje, če sam zapis nikoli ne uide? Ker je *članstvo dejstvo*. Če je bil model učen na bolnikih, ki so prejemali določeno imunoterapijo proti raku, lahko potrditev, da je vaš zapis v njem, razkrije, da ste verjetno imeli to bolezen — informacijo, ki je zavarovalnica ali delodajalec ne bi smela sklepati. Vsebina ostane zaprta; uide dejstvo pripadnosti.
Avtorji predpostavke jasno naštejejo — dostop do običajnih napovedi modela, kandidatni zapis in lastni referenčni model napadalca — ne kot navodilo za napad, temveč zato, da lahko grožnjo ocenjujemo konkretno.

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.

Napad ne potrebuje posebnega vmesnika za zasebnost. Navadna diagnostična poizvedba lahko razkrije signal članstva, če je model nenavadno samozavesten pri zapisu, ki ga je že videl.

Kaj so ugotovili

Tri ugotovitve, vsaka ostrejša od prejšnje.

Povprečja skrijejo izpostavljene. Agregatno — kot se običajno meri — je veliko teh modelov videti pomirjujoče zasebnih: pri večini bolnikov napad ni nič boljši od naključja. Pogled po posameznikih pa pokaže drugo sliko. Kot je Knolle povedal v objavi študije, so prejšnje ocene »merile le povprečno tveganje vseh bolnikov. Mi smo ga prvič pregledali na ravni posameznih bolnikov — in slika je zelo drugačna.« Pri nekaterih posameznikih napad uspe **skoraj popolnoma** — AUC napada je **0,95 ali več** (0,5 pomeni met kovanca, 1,0 popolno ločevanje) — čeprav je povprečje celotnega nabora videti kot naključje.

Povprečje je lahko blizu naključju, medtem ko majhen rep zapisov ostane bistveno bolj izpostavljen. Bistvo članka je, da je treba zasebnost preverjati na ravni bolnika, ne le agregatno.

Original Aurora diagram — The Clean Paper CC BY 4.0

Izpostavljenost ni enaka — najbolj prizadene podzastopane. Bolniki, ki so bili najbolj ranljivi za skoraj popoln napad, so sistematično prihajali iz skupin, **premalo zastopanih v podatkih**: manjšinskih etničnih skupin, redkih bolezenskih fenotipov, nenavadnih slikovnih značilnosti. V zbirki elektronskih zapisov so se temnopolti bolniki med najbolj ranljivimi zapisi pojavljali **31 % pogosteje**, kot bi pričakovali; v mamografski zbirki so bile slike, označene kot sumljive za malignost, v tem skrajnem repu prekomerno zastopane za **+1.179 %**. (To sta primera, ne celotna slika — članek kaže enako smer pri več skupinah — in številki pomenita *relativno* prekomerno zastopanost v majhnem repu skrajnega tveganja, ne deleža vseh temnopoltih bolnikov ali sumljivih mamografij, ki so izpostavljeni.) Model ima manj podobnih primerov, v katere bi se ti bolniki »zlili«, zato bolj izstopajo — in prav izstopanje je

tisto, kar napad zazna. Odpoved zasebnosti ni naključna; koncentrira se na ljudeh na robu podatkov.

Večji modeli tveganje povečajo. Število bolnikov, izpostavljenih skoraj popolnim napadom, je z zmogljivostjo modela močno naraščalo — v eni dermatološki zbirki se je delež povzpel od praktično nič pri najmanjšem modelu do približno **enega od desetih** pri največjem. Ker medicinski modeli postajajo večji in zmogljivejši — smer celotnega področja — se prav to tveganje po opozorilu avtorjev povečuje, ne zmanjšuje.

Rückertov povzetek je neposreden: »To ni sprejemljivo tveganje. Zdravstveni podatki so izjemno občutljivi.«

Zakaj je »varno v povprečju« napačen preizkus

Nizko povprečno tveganje je mamljivo brati kot potrdilo o varnosti. Osrednja lekcija članka je, da povprečenje tu ni le nenatančno — meri napačno stvar.

Jamstvo zasebnosti ima pomen le, če velja za človeka z največjim tveganjem, ne za medianega človeka. Model, pri katerem 999 od 1.000 bolnikov ni mogoče prepoznati, enega pa skoraj popolnoma, ni »99,9-odstotno zaseben« v nobenem smislu, ki bi temu enemu človeku kaj pomenil. Če je ta človek predvidljivo bolnik z redko boleznijo ali pripadnik etnične manjšine, metrika ni samo nepopolna, temveč tiho diskriminatorna: izmeri varnost večine in jo razglasi za varnost vseh.

To je premik, ki ga članek zahteva: od vprašanja *kako zaseben je ta model v povprečju?* k *kdo je najbolj izpostavljen bolnik in kdo je ta oseba?* To sta različni vprašanji in le drugo je dejansko vprašanje zasebnosti.

Česa to ne dokazuje — in česa avtorji ne trdijo

- **Ne** pravi, da je treba medicinsko AI opustiti. Okvir avtorjev je ublažitev tveganja, ne umik: tveganje meriti in popravljati, ne prenehati graditi.
- **Ne** pomeni, da vsak medicinski model uhaja ali da je bil katerikoli določen uvedeni model napaden. Kaže, da *tveganje obstaja in ni enakomerno porazdeljeno* ter da ga standardne agregatne metrike spregledajo.
- **Ne** pomeni, da so vaši zapisi že razkriti. Napad zahteva posebne pogoje — dostop do modela, kandidatni zapis in napadalčevo infrastrukturo — ne zgolj naključno možnost.
- **Ne** kaže, da uhaja *vsebina* bolnikovega zapisa. Uhaja članstvo — dejstvo vključitve — ki je nevarno iz drugega razloga, ne zato, ker model izpiše kartoteko.
- **Ne** skrči neenakosti na eno samo naslovno številko. Močan in ponovljiv rezultat je *vzorec*: agregatne metrike podcenijo individualno tveganje, največ ga nosijo podzastopani bolniki, in to se ponavlja med zbirkami in stotinami modelov.

Kako močni so dokazi?

Gre za empiričen metodološki rezultat in je robusten. Vzorec — agregatne metrike zasebnosti sistematično podcenjujejo tveganje posameznika, preostalo tveganje se koncentrira na podzastopanih skupinah in narašča z zmogljivostjo modela — se je pojavil v **sedmih zbirkah različnih tipov podatkov** in pri velikem številu modelov na zbirko. Prav ta širina naredi merilno trditev verodostojnejšo od artefakta ene zbirke.

Dve pošteni omejitvi. Prvič, to je prikaz *tveganja*, izmerjen z napadi, ki so jih avtorji sami izvedli; opisuje, kako dobro bi sposoben napadalec *lahko* deloval pod navedenimi predpostavkami, ne kako pogosto se resnični napadi dogajajo. Drugič, najbolj dramatične številke — skoraj popoln uspeh napada pri nekaterih bolnikih — po zasnovi opisujejo najslabše zaščitene posameznike. To je bistvo analize in treba jih je brati kot »rep je mnogo težji, kot nakazuje povprečje«, ne kot »večina bolnikov je izpostavljena«.

Primeren odziv ni alarmizem niti odmahovanje: gre za skrbno večpodatkovno študijo, ki kaže, da je standardni način potrjevanja zasebnosti medicinske AI slep za lastne najslabše primere — in da ti primeri padejo na ljudi z najmanj rezerve.

Zakaj je pomembno

Dve stvari to naredita več kot tehnično opombo.

Prvič, spreminja, kaj bi smelo pomeniti »varovanje zasebnosti«. Model je lahko izdan z resnično nizko povprečno oceno tveganja, pa še vedno skriva podskupino bolnikov, ki jih napad članstva skoraj popolnoma prepozna. Praktična zahteva članka je konkretna: pred izdajo oceniti tveganje **za vsakega bolnika**, nadzorovati dostop do uvedenih modelov in uporabljati tehnike, kot je **diferencialna zasebnost** — majhen, skrbno umerjen šum med učenjem, ki otopi napade članstva, ob resničnem in upravljanem strošku za uporabnost modela. Kompromis med zasebnostjo in uporabnostjo je izrecen, ne brezplačen. »Preverili smo povprečje« ne bi smelo več šteti kot dovolj.

Drugič, študija poveže zasebnost s pravičnostjo. Iste skupine, ki so premalo zastopane v medicinskih podatkih — in jih zato medicinska AI že slabše obravnava — so tudi skupine, katerih zasebnost ta AI najslabše ščiti. Področje, ki poskuša reševati prvo neenakost, druge ne more odložiti na nekoga drugega. Gre za iste ljudi.

Pomirjujoča zgodba o zasebnosti medicinske AI — *izmerili smo jo, povprečje je nizko, vse je v redu* — je zgodba, ki jo ta članek razstavi. Ne zato, da bi ljudi odvrnil od tehnologije, temveč da bi standard postavil tja, kamor sodi: obljubo lahko trdimo, da držimo, šele ko smo jo preverili pri človeku, ki jo najlažje izgubi.

Kratek povzetek

Napadi sklepanja o članstvu poskušajo ugotoviti, ali je bil zapis določene osebe v učnem naboru modela AI — in ker je že članstvo lahko razkrivajoče, denimo da ste imeli določeno bolezen, je to resnično uhajanje zasebnosti tudi brez razkritja same kartoteke. Prejšnje raziskave so merile, kako pogosto taki napadi uspejo *v povprečju* čez podatkovni nabor. Ta študija je tveganje merila *po bolniku* na sedmih medicinskih zbirkah (slikanje, EKG, elektronski zapisi) in številnih modelih ter pokazala, da agregatne metrike tveganje močno podcenijo: nekatere posameznike je mogoče prepoznati skoraj popolnoma (AUC napada $\geq$ 0,95), čeprav je povprečje videti varno; najranljivejši sistematično prihajajo iz podzastopanih skupin (manjšinske etnične skupine, redke bolezni, nenavadne slikovne značilnosti); problem pa narašča z velikostjo modela. Avtorji ne pozivajo k opustitvi medicinske AI — zahtevajo merjenje zasebnosti na ravni posameznika, nadzor dostopa in diferencialno zasebnost. Sklep: »zasebno v povprečju« ni jamstvo zasebnosti, odpove pa navadno prav ljudem, ki so že najmanj zaščiteni.

Preverjanje brez olepševanja

Kaj članek pokaže: V sedmih medicinskih podatkovnih zbirkah in številnih modelih je tveganje napada članstva pri nekaterih posameznikih precej višje, kot nakazujejo agregatne metrike — vse do skoraj popolnega uspeha napada (AUC $\geq$ 0,95) — preostalo tveganje pa nesorazmerno prizadene podzastopane skupine in narašča z zmogljivostjo modela.

Kaj je verjetno, vendar ne dokazano: Da resnični napadalci to že izkoriščajo proti uvedenim kliničnim modelom. Študija pokaže *zmožnost in njeno porazdelitev* pod določenimi predpostavkami, ne pogostosti napadov v naravi.

Česa ne pokaže: Da bi morali medicinsko AI opustiti; da vsak model uhaja ali da je bil določen uvedeni model vdrt; da uhaja *vsebina* zdravstvene kartoteke in ne članstvo; ene same številke, ki bi povzela vse neenakosti — robusten rezultat je vzorec, ne ena vrednost.

Glavne omejitve: Meri najslabše tveganje z napadi, ki so jih izvedli avtorji, ne dejanskih incidentov; dramatične številke po zasnovi opisujejo najbolj izpostavljene posameznike; točne razlike med skupinami so predstavljene kot konsistenten vzorec čez podatkovne zbirke, ne kot ena univerzalna številka.

Koliko zaupanja naj ima splošni bralec? Visoko, da agregatne metrike zasebnosti podcenjujejo tveganje posameznika, da se preostalo tveganje koncentrira na podzastopanih bolnikih in da se povečuje z velikostjo modela — vse troje je prikazano na številnih zbirkah in modelih. Zmerno glede pogostosti takšnih napadov v resničnem svetu, česar študija ne meri. Varno branje ni »medicinska AI razkriva vaše podatke« in ne »zasebnost je rešena«, temveč »standardni preizkus zasebnosti je slep za lastne najslabše primere in ti padejo na najranljivejše bolnike — zato se mora preizkus spremeniti«.

Viri

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