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

Medicīniskais MI var būt privāts vidēji, bet tomēr atklāt konkrētus pacientus — visvairāk nepietiekami pārstāvētās grupas

Medicīniska MI modeļa apzīmējums "privātumu saglabājošs" parasti balstās vienā vidējā rādītājā. Šis pētījums apgalvo, ka tas ir nepareizais kritērijs. Analizējot dalības noteikšanas uzbrukumus — mēģinājumus noskaidrot, vai konkrētas personas ieraksts bija modeļa mācību datos un tādējādi netieši atklāt, piemēram, noteiktas slimības faktu — autori risku mērīja katram pacientam atsevišķi, nevis tikai kopumā. Septiņās medicīnas datu kopās (attēli, EKG, elektroniskie veselības ieraksti) un daudzos modeļos atkārtojās viena aina: modelis var izskatīties drošs vidēji, bet dažus cilvēkus uzbrucējs joprojām var identificēt gandrīz nevainojami (uzbrukuma AUC $\geq$ 0,95); visvairāk apdraudēti sistemātiski ir nepietiekami pārstāvēto grupu pacienti — etniskās minoritātes, retu slimību pacienti vai cilvēki ar neparastiem attēlveidošanas raksturlielumiem — un, modeļiem kļūstot lielākiem, problēma saasinās. Autori neaicina atteikties no medicīniskā MI: viņi iesaka privātumu vērtēt katram pacientam, kontrolēt piekļuvi modeļiem un izmantot diferenciālo privātumu. Neērtais kodols ir vienkāršs: "privāts vidēji" nav privātuma garantija, un visbiežāk tā pievilj tos pacientus, kuri jau tā ir vismazāk aizsargāti.

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

Privātuma garantija ir solījums cilvēkam, nevis vidējam rādītājam

Kad medicīnisku MI modeli sauc par "privātumu saglabājošu", šis apgalvojums parasti balstās vienā skaitlī: vidēji visiem pacientiem, kuru dati izmantoti modeļa apmācībā, iespēja noteikt, vai konkrēta persona bija mācību datu kopā, ir maza. Tas izklausās mierinoši. Šī pētījuma klusais un neērtais secinājums ir tāds, ka tiek skatīts nepareizais skaitlis.

Privātums nav vidējais rādītājs. Tas ir solījums katram cilvēkam atsevišķi — ka fakts par viņa datu izmantošanu apmācībā vēlāk nekļūs par veidu, kā viņu atklāt. Vidējais rādītājs var šo solījumu izpildīt gandrīz visiem, bet dažiem to pilnībā lauzt. Pētnieki privātuma risku novērtēja pacientu pa pacientam, ar līdz šim neizmantotu detalizāciju, un atrada tieši to: modeļi, kas kopumā izskatās droši, dažiem cilvēkiem var gandrīz nevainojami atklāt dalību mācību datos — un nesamērīgi bieži tie ir tie, kuri jau tā ir vismazāk aizsargāti.

Ko autori darīja

Komanda — Morica Knolles vadībā kopā ar Danielu Rikertu (abi Minhenes Tehniskajā universitātē) un Georgu Kaisisu (Hasso Plattner Institute), kā arī kolēģiem no Imperial College London — paņēma labi zināmu privātuma apdraudējumu, ko sauc par **dalības noteikšanas uzbrukumu** (*membership inference attack*), un mainīja jautājumu. Tā vietā, lai vaicātu "cik bieži šāds uzbrukums izdodas vidēji visā datu kopā?", viņi vaicāja: "cik liels ir risks tieši *šim pacientam*?"

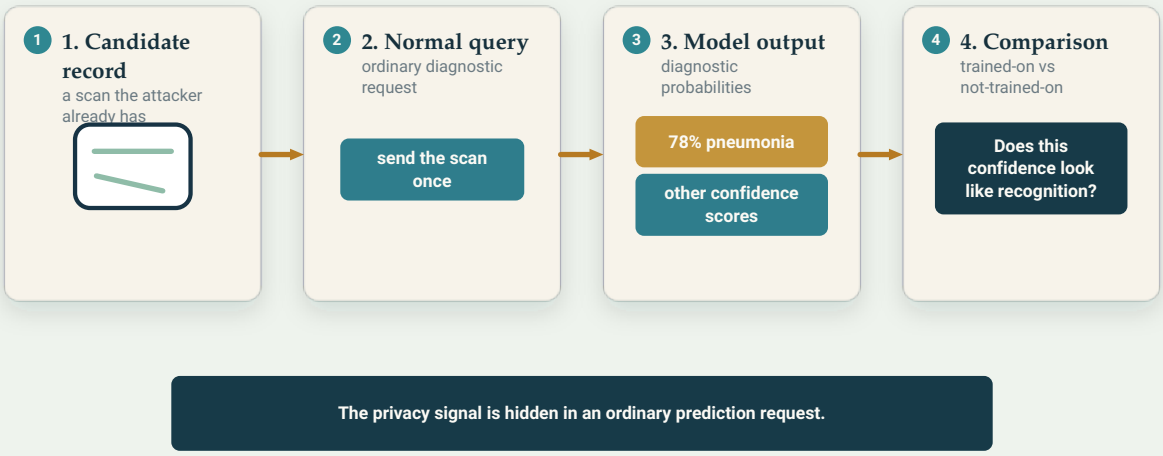
Šādu novērtējumu katram pacientam viņi veica **septiņās plaši izmantotās medicīnas datu kopās**, aptverot ļoti atšķirīgus datu veidus — medicīniskos attēlus, elektrokardiogrammas un elektroniskos veselības ierakstus — un katrā no tām analizēja 200 modeļus. Katram pacientam katrā modelī pētnieki novērtēja, cik droši uzbrucējs varētu pateikt, vai šīs personas ieraksts bija mācību datos, un pēc tam rezultātus sadalīja pa grupām: slimības statuss, paša norādītā rase, apdrošināšana, dzimums un attēlveidošanas protokols.

> **Kas patiesībā ir dalības noteikšanas uzbrukums**
>
> Noplūde šeit ir smalkāka par situāciju, kurā "modelis izvada jūsu medicīnisko ierakstu". Runa ir par *dalību* — pašu faktu, ka jūsu dati atradās mācību kopā.
> Sāksim ar to, ko šāds modelis dara ikdienā. Jūs tam iedodat, piemēram, krūškurvja rentgenu, un tas atdod varbūtības: 78% pneimonijas iespējamība, kā arī novērtējumi par kardiomegāliju, tūsku un konsolidāciju. Uzbrukums izmanto tikai šo parasto diagnostisko atbildi. Nekādas īpašas slepenas piekļuves nav.
> Vājā vieta ir šāda: modelis bieži ir nedaudz **pārliecinātāks** par tieši tiem piemēriem, kurus tas redzēja apmācībā, nekā par tādiem, kurus nekad nav redzējis. To var salīdzināt ar studentu, kurš eksāmena jautājumus jau slepus redzējis iepriekš — uz trenētajiem jautājumiem atbildes nāk mazliet pārāk ātri un pārāk droši. No šīs pazīmes mēģināt secināt, vai konkrētais ieraksts

bija viens no modeļa apmācības piemēriem, arī nozīmē “dalības noteikšana”.
Tātad uzbrukuma loģika ir šāda. Kāds grib noskaidrot, vai konkrētas personas attēls izmantots modeļa apmācībā. Viņš **neprasa** modelim izdot šo ierakstu — kandidāta attēls viņam jau ir, piemēram, pašas personas attēls vai ļoti tuva kopija. To vienreiz iesniedz kā parastu diagnostikas vaicājumu un fiksē modeļa pārliecību. Pēc tam šo pārliecības līmeni salīdzina ar to, kā parasti uzvedas modelis, kas konkrēto attēlu *ir* redzējis apmācībā, un modelis, kas to *nav* redzējis. Šādu atšķirību var apgūt ar paša uzbrucēja trenētu atsauces modeli uz parasta datora, bez īpašas aparatūras. Ja īstais modelis par konkrēto attēlu ir neparasti pārliecināts — it kā to atpazītu — tas ir spēcīgs signāls, ka attēls bijis mācību datos.
Satraucošais ir tas, ka pats vaicājums nemaz neizskatās pēc uzbrukuma: tas ir tāds pats diagnostikas pieprasījums, kādu nosūtītu ārsts, un privātuma signāls slēpjas parastā prognozē. Tieši tāpēc risks ir konkrēts, lai gan līdz šim tas parādīts definētos laboratorijas apstākļos, nevis dokumentēts reālā uzbrukumā.
Kāpēc dalība vispār ir svarīga, ja pats ieraksts nekad netiek atklāts? Tāpēc, ka *dalība pati par sevi ir fakts*. Ja modelis apmācīts, piemēram, ar pacientiem, kuri saņēmuši noteiktu vēža imūnterapiju, apstiprinājums, ka jūsu ieraksts ir šajā kopā, var netieši atklāt, ka jums, visticamāk, bijis šis vēzis — tieši tādu informāciju apdrošinātājam vai darba devējam nevajadzētu spēt izsecināt. Ieraksta saturs paliek noslēpts; ārā izkļūst fakts par piederību kopai.
Autori šos priekšnoteikumus izklāsta skaidri — piekļuve modeļa parastajām prognozēm, kandidāta ieraksts un paša uzbrucēja atsauces modelis — nevis kā praktisku pamācību, bet lai par apdraudējumu varētu spriest konkrēti, nevis tikai abstrakti.

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.



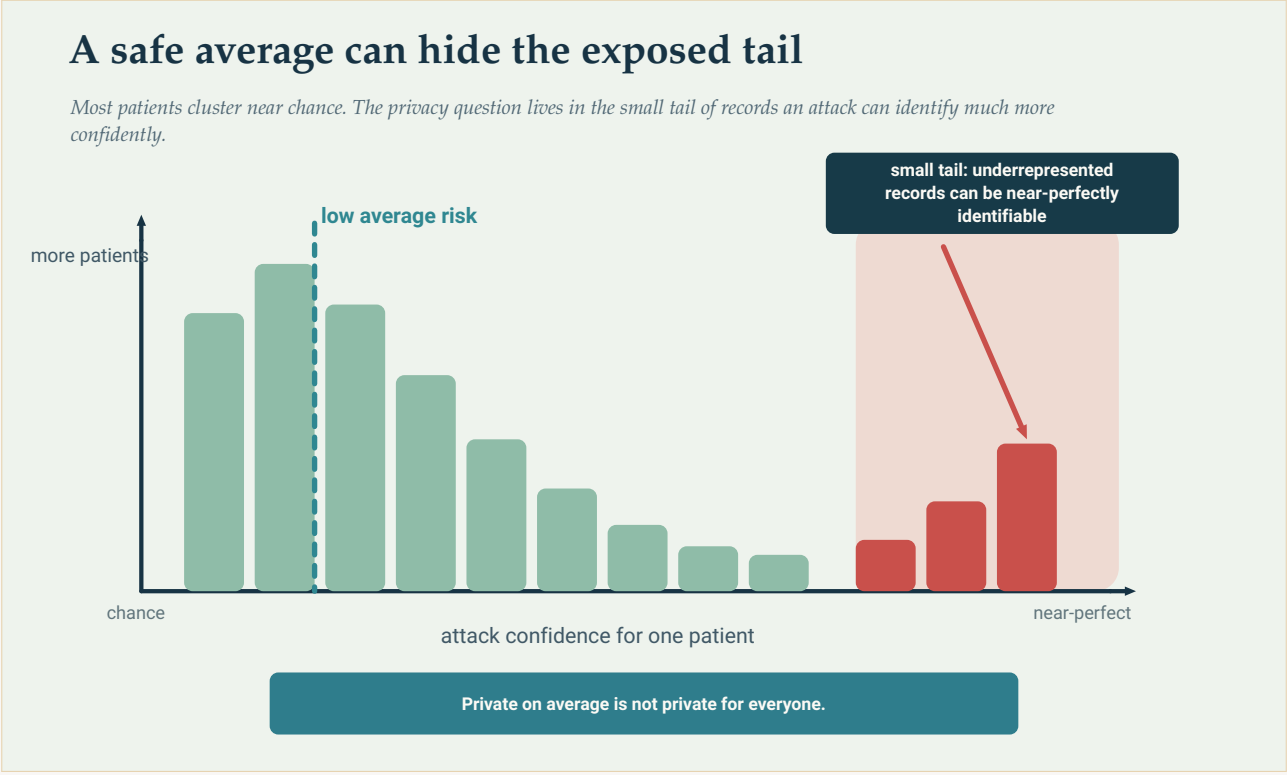
Mechanism only: no pseudocode, no attack threshold, no recipe.

Uzbrukumam nav vajadzīga īpaša privātuma API. Parasts diagnostikas vaicājums var saturēt signālu par dalību mācību datos, ja modelis par iepriekš redzētu ierakstu ir neparasti pārliecināts.

Ko viņi atklāja

Trīs rezultāti, un katrs nākamais ir neērtāks par iepriekšējo.

Vidējais rādītājs paslēpj visvairāk apdraudētos. Ja privātumu mēra kopumā — kā tas parasti tiek darīts — daudzi modeļi izskatās nomierinoši droši: lielākajai daļai pacientu uzbrukums nav labāks par nejaušu minējumu. Pacientu līmeņa analīze parādīja citu ainu. Kā Knolle formulēja pētījuma paziņojumā, iepriekšējie novērtējumi "vienmēr mērījuši tikai vidējo risku visiem pacientiem. Mēs pirmo reizi aplūkojām risku atsevišķu pacientu līmenī — un aina ir ļoti atšķirīga." Dažiem cilvēkiem uzbrukums izdodas **gandrīz nevainojami**: autori to definē kā uzbrukuma AUC **0,95 vai vairāk** (0,5 atbilst monētas mešanai, 1,0 — nevainojamai atšķiršanai), pat ja visas datu kopas vidējais rezultāts nav labāks par nejaušību.



Vidējais rādītājs var atrasties tuvu nejaušības līmenim, lai gan neliela ierakstu grupa joprojām ir daudz vieglāk identificējama. Pētījuma galvenais vēstījums: privātums jāpārbauda pacienta līmenī, ne tikai kopumā.

Risks nav sadalīts vienādi — tas koncentrējas nepietiekami pārstāvētajās grupās. Pacienti, kuri bija visneaizsargātākie pret gandrīz nevainojamu uzbrukumu, sistemātiski piederēja grupām, kas datos bija **nepietiekami pārstāvētas**: etniskajām minoritātēm, retu slimību fenotipiem, neparastiem attēlveidošanas raksturlielumiem. Elektronisko veselības ierakstu datu kopā melnādainie pacienti visaugstākā riska ierakstu grupā parādījās **par 31% biežāk**, nekā būtu sagaidāms; mammogrāfijas

datu kopā attēli, kas bija atzīmēti kā aizdomīgi attiecībā uz ļaundabīgu procesu, šajā ārkārtīgi augsta riska grupā bija pārstāvēti par iespaidīgiem **+1179%** vairāk. (Tie ir divi piemēri, ne visa aina — līdzīgs nobīdes modelis pētījumā redzams vairākās grupās. Turklāt tie ir *relatīvi pārstāvētības pieaugumi* nelielajā galējā riska "astē": tie norāda, cik daudz biežāk konkrētie pacienti nonāk visvairāk apdraudēto ierakstu vidū, nevis cik liela daļa no visiem melnādainajiem pacientiem vai aizdomīgajām mammogrāfijām ir apdraudēta.) Modelim ir mazāk līdzīgu piemēru, kuros šādu pacientu "izšķīdināt", tāpēc viņu ieraksti izceļas — un tieši izcelšanos dalības noteikšanas uzbrukums izmanto. Privātuma kļūme nav nejauša; tā koncentrējas cilvēkos, kuri jau tā atrodas malā.

Lielāki modeļi problēmu pastiprina. Pacientu skaits, kuriem uzbrukums bija gandrīz nevainojams, strauji pieauga līdz ar modeļa ietilpību. Vienā dermatoloģijas datu kopā šādu pacientu īpatsvars pieauga no praktiski nulles mazākajā modelī līdz aptuveni **vienam no desmit** lielākajā. Medicīniskajiem MI modeļiem kļūstot arvien lielākiem un spējīgākiem — tieši tādā virzienā nozare pašlaik attīstās — šis konkrētais risks, pēc autoru brīdinājuma, nevis mazinās, bet pieaug.

Rikerta kopsavilkums ir tiešs: "Tas nav pieņemams risks. Veselības dati ir ārkārtīgi sensitīvi."

Kāpēc "drošs vidēji" ir nepareizais kritērijs

Ir vilinoši zemu vidējo risku uztvert kā apliecinājumu, ka viss ir kārtībā. Pētījuma galvenā doma ir tāda, ka vidējais rādītājs šeit nav tikai neprecīzs — tas mēra nepareizo lietu.

Privātuma garantijai ir jēga tikai tad, ja tā attiecas arī uz cilvēku ar vislielāko risku, nevis tikai uz tipisko pacientu. Modeli, kurā 999 no 1000 pacientiem nav iespējams atšķirt, bet vienu var identificēt gandrīz nevainojami, nevar jēgpilni saukt par "99,9% privātu" attiecībā uz šo vienu cilvēku. Ja turklāt šis viens cilvēks paredzami ir retas slimības pacients vai etniskās minoritātes pārstāvis, metrika nav vienkārši nepilnīga — tā klusām rada nevienlīdzību. Tā izmēra drošību vairākumam un nosauc to par drošību visiem.

Tieši šo jautājuma maiņu pētījums uzspiež: no *cik privāts šis modelis ir vidēji?* uz *kurš pacients ir visvairāk apdraudēts, un kas viņš ir?* Tie ir dažādi jautājumi, un tikai otrais patiešām ir jautājums par privātumu.

Ko tas *nepierāda* un neapgalvo

- Tas **nenozīmē**, ka no medicīniskā MI būtu jāatsakās. Autoru pieeja ir risku mazināt, nevis atkāpties: izmērīt, novērst un turpināt būvēt.
- Tas **nenozīmē**, ka noplūst katrs medicīniskais modelis vai ka kāds konkrēts darbībā esošs modelis jau ir uzlauzts. Pētījums parāda, ka *risks pastāv un nav sadalīts vienādi,* un ka standarta kopējie rādītāji to neuzrāda.
- Tas **nenozīmē**, ka jūsu medicīniskie ieraksti jau ir atklāti. Uzbrukumam vajadzīgi konkrēti priekšnoteikumi — piekļuve modelim, kandidāta ieraksts un paša uzbrucēja infrastruktūra — tā

nav nejauši pieejama spēja.

- Tas **neparāda**, ka noplūst pacienta ieraksta *saturs*. Noplūst dalība — fakts par iekļaušanu mācību datos — un tas ir bīstami cita iemesla dēļ, nevis tāpēc, ka modelis izdotu visu ierakstu.
- Tas **nesavelk** nevienlīdzību vienā skaļā skaitlī. Spēcīgais un atkārtojamais rezultāts ir *modelis*: kopējie rādītāji nenovērtē individuālo risku, un vislielāko slogu nes nepietiekami pārstāvētie pacienti — tas atkārtojas dažādās datu kopās un simtos modeļu.

Cik pārliecinoši ir pierādījumi?

Tas ir empīrisks metodoloģisks rezultāts, un tas ir robusts. Aina — kopējās privātuma metrikas sistemātiski nenovērtē risku atsevišķiem pacientiem, atlikušais risks koncentrējas nepietiekami pārstāvētajās grupās un pieaug līdz ar modeļa ietilpību — atkārtojās **septiņās dažādu tipu datu kopās** un daudzos modeļos katrā no tām. Tieši šis plašums padara mērījuma secinājumu ticamāku nekā viena datu kopas īpatnību.

Ir divi svarīgi ierobežojumi. Pirmkārt, pētījums demonstrē *risku*, izmantojot autoru pašu īstenotus uzbrukumus; tas raksturo, cik labi spējīgs uzbrucējs *varētu* darboties pie noteiktiem priekšnoteikumiem, nevis cik bieži šādi uzbrukumi notiek reālajā pasaulē. Otrkārt, iespaidīgie skaitļi — gandrīz nevainojama atšķiršana dažiem pacientiem — pēc definīcijas attiecas uz visvairāk apdraudētajiem indivīdiem. Tā arī ir pētījuma doma. Tos jālasa kā "riska sadalījuma aste ir daudz smagāka, nekā rāda vidējais", nevis kā "lielākā daļa pacientu ir atklāti".

Tāpēc pareizā reakcija nav ne panika, ne noraidījums. Tas ir rūpīgs vairāku datu kopu pētījums, kas rāda, ka parastais medicīniskā MI privātuma novērtēšanas veids neredz savus sliktākos gadījumus — un ka tieši šie sliktākie gadījumi skar pacientus, kuriem jau tā ir vismazāk aizsardzības rezervju.

Kāpēc tas ir svarīgi

Divi aspekti padara šo darbu par ko vairāk nekā tehnisku piezīmi.

Pirmkārt, tas maina to, ko vispār drīkstētu saukt par "privātumu saglabājošu". Modelim var būt patiesi zems vidējais privātuma risks un tomēr neliela pacientu grupa, kuru dalību mācību datos iespējams noteikt gandrīz nevainojami. Pētījuma praktiskā prasība ir konkrēta: pirms modeļa izlaišanas novērtēt privātuma risku **katram pacientam**, kontrolēt piekļuvi izvietotajiem modeļiem un izmantot tādas metodes kā **diferenciālais privātums** — nelielu, rūpīgi kalibrētu troksni apmācības laikā, kas mazina dalības noteikšanas uzbrukumu efektivitāti, bet par reālu un pārvaldāmu cenu modeļa lietderībai. Privātuma un lietderības kompromiss ir atklāts; tas nav bezmaksas risinājums. Ar frāzi "mēs pārbaudījām vidējo" vairs nevajadzētu pietikt.

Otrkārt, pētījums sasaista privātumu ar taisnīgumu. Tās pašas grupas, kas medicīnas datos ir nepietiekami pārstāvētas — un tāpēc medicīniskais MI jau var tās apkalpot sliktāk — izrādās arī tās, kuru privātumu šis MI aizsargā visvājāk. Nozare, kas nopietni strādā pie pirmās nevienlīdzības, nevar

izlikties, ka otrā pieder citai problēmu kategorijai. Tie ir tie paši cilvēki.

Mierinošais stāsts par medicīniskā MI privātumu — *mēs to izmērījām, vidējais ir zems, tātad viss kārtībā* — ir tieši tas, ko šis pētījums izjauc. Nevis lai kādu atbaidītu no tehnoloģijas, bet lai standartu pārvietotu tur, kur tam jābūt: privātuma solījumu drīkst apgalvot tikai tad, ja esi pārbaudījis, vai tas turas arī attiecībā uz cilvēku, kurš visdrīzāk var ciest.

Īss kopsavilkums

Dalības noteikšanas uzbrukumi mēģina noskaidrot, vai konkrētas personas ieraksts bija MI modeļa mācību datos. Tā kā pats dalības fakts var atklāt sensitīvu informāciju — piemēram, ka cilvēkam bijusi noteikta slimība — tā ir īsta privātuma noplūde pat tad, ja medicīniskā ieraksta saturs nekad neparādās. Iepriekšējos darbos parasti mērīja, cik bieži šādi uzbrukumi izdodas *vidēji* visā datu kopā. Šis pētījums risku mērīja *katram pacientam* septiņās medicīnas datu kopās — attēlveidošanā, EKG un elektroniskajos veselības ierakstos — un daudzos modeļos katrā no tām. Rezultāts: kopējās metrikas risku būtiski nenovērtē. Dažus cilvēkus var identificēt gandrīz nevainojami (uzbrukuma AUC $\geq$ 0,95), pat ja visas datu kopas vidējais rādītājs izskatās drošs; visneaizsargātākie sistemātiski ir nepietiekami pārstāvēto grupu pacienti — etniskās minoritātes, retu slimību pacienti, cilvēki ar neparastiem attēlveidošanas raksturlielumiem — un problēma pieaug līdz ar modeļa izmēru. Autori neaicina atteikties no medicīniskā MI: viņi aicina privātumu mērīt individuālā līmenī, kontrolēt piekļuvi modeļiem un izmantot diferenciālo privātumu. Galvenais secinājums: "privāts vidēji" nav privātuma garantija, un visbiežāk tā pieviļ tos, kuri jau tā ir vismazāk aizsargāti.

Bez pārspīlējumiem

Ko pētījums parāda: Septiņās medicīnas datu kopās un daudzos modeļos katrā no tām dalības noteikšanas risks atsevišķiem pacientiem ir daudz augstāks, nekā liek domāt kopējās metrikas — līdz pat gandrīz nevainojamai atšķiršanai (AUC $\geq$ 0,95). Šis atlikušais risks nesamērīgi skar nepietiekami pārstāvētas grupas un pieaug līdz ar modeļa ietilpību.

Kas ir ticams, bet nav pierādīts: Ka reāli uzbrucēji to jau sistemātiski izmanto pret darbībā esošiem klīniskajiem modeļiem. Pētījums demonstrē *spēju un tās sadalījumu* pie konkrētiem priekšnoteikumiem, nevis reālo uzbrukumu biežumu.

Ko tas neparāda: Ka no medicīniskā MI jāatsakās; ka noplūst katrs modelis vai ka konkrēts darbībā esošs modelis jau ir kompromitēts; ka tiek atklāts pacienta ieraksta *saturs*, nevis dalības fakts; vai ka nevienlīdzību var raksturot ar vienu universālu skaitli — robusts ir pats atkārtojošais modelis, nevis viens procents.

Galvenie ierobežojumi: Tiek mērīts sliktākā gadījuma risks ar autoru pašu īstenotiem uzbrukumiem, nevis dokumentēti reāli incidenti; dramatiskākie skaitļi pēc definīcijas raksturo visvairāk apdraudētos indivīdus; un precīzās atšķirības starp grupām tiek parādītas kā konsekvents modelis vairākās

datu kopās, nevis reducētas uz vienu skaitli.

Cik lielai pārliecībai vajadzētu būt vispārīgam lasītājam? Augstai attiecībā uz to, ka kopējās privātuma metrikas nenovērtē individuālo risku, ka atlikušais risks koncentrējas nepietiekami pārstāvētajos pacientos un ka tas pieaug līdz ar modeļa izmēru — tas parādīts daudzās datu kopās un modeļos. Mērenai attiecībā uz šādu uzbrukumu biežumu reālajā pasaulē, ko šis pētījums nemēra. Drošā interpretācija nav "medicīniskais MI nopludina jūsu datus" un nav arī "privātuma problēma ir atrisināta", bet gan: "standarta privātuma pārbaude neredz savus sliktākos gadījumus, un tie skar visneaizsargātākos pacientus — tāpēc pārbaude ir jāmaina."

Avoti

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