

The Clean Paper

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

Medicininis DI gali būti privatus vidutiniškai ir vis tiek atskleisti konkrečius pacientus – ypač nepakankamai atstovaujamus

Medicininio DI modelio „privatumas“ dažnai apibūdinamas vienu vidutiniu skaičiumi. Šis tyrimas teigia, kad tai netinkamas testas. Autoriai nagrinėjo narystės nustatymo atakas – metodus, leidžiančius spręsti, ar konkretaus žmogaus įrašas buvo modelio mokymo duomenyse ir taip netiesiogiai atskleisti jautrią informaciją – ir riziką matavo kiekvienam pacientui atskirai septyniuose medicinos duomenų rinkiniuose bei daugybėje modelių. Rezultatas: modeliai, kurie vidutiniškai atrodo saugūs, kai kuriuos žmones vis tiek leidžia nustatyti beveik idealiai (atakos AUC $\geq$ 0,95); labiausiai pažeidžiami sistemingai yra nepakankamai atstovaujamų grupių pacientai; o augant modeliams problema stiprėja. Autoriai nesiūlo atsisakyti medicininio DI – jie siūlo vertinti privatumą paciento lygmeniu, kontroliuoti prieigą ir taikyti diferencinį privatumą. Esmė nepatogi: „privatus vidutiniškai“ nėra privatumo garantija.

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

Privatumo garantija yra pažadas žmogui, o ne vidurkiui

Kai medicininio DI modelis vadinamas „sauganˇciu privatumą“, šis teiginys paprastai remiasi vienu skaičiumi: vertinant visus pacientus, kurių duomenys naudoti mokymui, vidutinė tikimybė nustatyti konkretaus žmogaus dalyvavimą mokymo rinkinyje yra maža. Skamba raminamai. Tylus ir nepatogus šio darbo teiginys – tai ne tas skaičius, kurį turėtume vertinti.

Privatumas nėra vidurkis. Tai pažadas kiekvienam žmogui, kad jo patekimas į mokymo rinkinį vėliau jo neatskleis. O vidurkis gali šį pažadą ištesėti beveik visiems ir visiškai sulaužyti keliems. Tyrėjai privatumo riziką išmatavo kiekvienam pacientui atskirai – tokiu detalumu, kokio ankstesniuose darbuose nebuvo – ir rado būtent tai: modeliai, kurie bendruose rodikliuose atrodo saugūs, kai kurių konkrečių žmonių narystę mokymo rinkinyje gali atskleisti beveik nepriekaištingai. Ir neproporcingai dažnai tai nutinka tiems, kurie jau ir taip yra prasčiausiai atstovaujami.

Ką padarė autoriai

Moritz Knolle vadovaujama komanda, kurioje dirbo Daniel Rückert (abu Miuncheno technikos universitete), Georg Kaissis (Hasso Plattner institutas) ir kolegos iš Londono imperatoriškojo koledžo, paėmė gerai žinomą privatumo grėsmę – **narystės nustatymo ataką** (*membership inference attack*) – ir pakeitė klausimą. Užuot klausę „kaip dažnai ši ataka vidutiniškai pavyksta visame duomenų rinkinyje?“, jie klausė: „kiek pažeidžiamas yra *šis konkretus pacientas*?“

Tokį kiekvieno paciento vertinimą jie atliko **septyniuose plačiai naudojamuose medicinos duomenų rinkiniuose**, apimančiuose labai skirtingus duomenų tipus – medicininius vaizdus, elektrokardiogramas ir elektroninius sveikatos įrašus – ir kiekviename jų ištyrė po 200 modelių. Kiekvienam pacientui kiekviename modelyje tyrėjai apskaičiavo, kaip užtikrintai užpuolikas galėtų nustatyti, ar to žmogaus įrašas buvo mokymo duomenyse. Tada riziką suskirstė pagal grupes: ligos būklę, paties paciento nurodytą rasę, draudimą, lytį ir vaizdinimo protokolą.

> **Kas iš tikrųjų yra narystės nustatymo ataka?**
>
> Čia nutekėjimas subtilesnis nei „modelis išspjauna jūsų medicinos įrašą“. Nuteka *narystė* – pats faktas, kad jūsų duomenys buvo mokymo rinkinyje.
> Pradėkime nuo įprasto modelio darbo. Jam pateikiama, tarkime, krūtinės rentgenograma, o jis grąžina tikimybes: 78 % pneumonijos tikimybė, taip pat kardiomegalijos, edemos ar konsolidacijos įverčiai. Atakai naudojamas tik šis įprastas diagnostinis atsakymas. Jokios specialios prieigos nereikia.
> Silpnybė, kuria remiamasi, tokia: modelis dažnai būna šiek tiek **labiau užtikrintas** dėl tų konkrečių pavyzdžių, kuriuos matė mokymo metu, nei dėl niekada nematytų. Galima įsivaizduoti studentą, kuris slapta iš anksto matė egzamino klausimus – prie išmoktų klausimų atsakymai ateina truputį per greitai ir per užtikrintai. Nustatyti, ar konkretus įrašas buvo vienas iš modelio „mokytų“ pavyzdžių, būtent ir reiškia narystės nustatymą.

Ataka vyksta maždaug taip. Kažkas nori sužinoti, ar konkretaus žmogaus vaizdas buvo naudotas modelio mokymui. Užpuolikas **neprašo** modelio atiduoti įrašo – kandidatą, pavyzdžiui, to žmogaus skenogramą ar labai artimą kopiją, jis jau turi. Vaizdas vieną kartą pateikiamas kaip visiškai įprasta diagnostinė užklausa ir užfiksuojamas modelio užtikrintumas. Tada tikrinama, ar šis užtikrintumas labiau panašus į modelio, kuris *mokėsi* iš to vaizdo, ar į modelio, kuris jo *nematė*. Tokį palyginimą galima išmokti susikūrus nuosavą pakaitinį „referencinį modelį“, net ir paprastu kompiuteriu. Jei tikrasis modelis dėl konkretaus vaizdo neįprastai užtikrintas – tarsi jį atpažintų – tai tampa stipriu ženklu, kad vaizdas buvo mokymo rinkinyje.

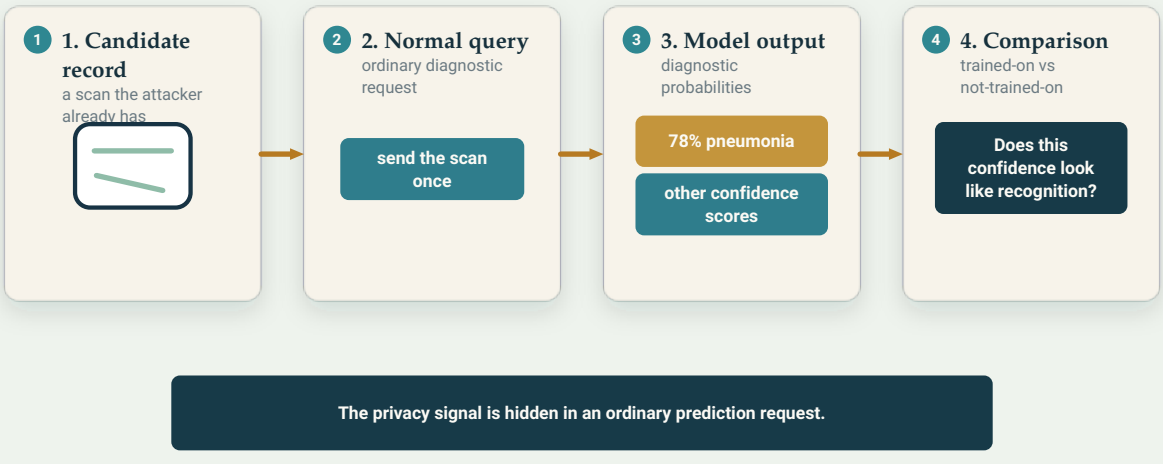
Neramina tai, kad pati užklausa visiškai neatrodo kaip ataka: ji tokia pati, kokią pateiktų klinicistas, o privatumo signalas slepiasi įprastoje prognozėje. Dėl to rizika konkreti, nors iki šiol ji parodyta nurodytomis laboratorinėmis sąlygomis, o ne užfiksuota realiame incidente.

Kodėl narystė svarbi, jei pats įrašas niekada nepasirodo? Nes *narystė yra faktas*. Jei modelis buvo mokytas iš pacientų, gavusių tam tikrą vėžio imunoterapiją, patvirtinimas, kad jūsų įrašas buvo mokymo rinkinyje, gali atskleisti, kad tikriausiai sirgote tuo vėžiu – informacija, kurios draudikas ar darbdavys neturėtų galėti išvesti. Įrašo turinys lieka uždaras; nuteka pats priklausymo rinkiniui faktas.

Autoriai aiškiai nurodo atakos prielaidas – prieigą prie įprastų modelio prognozių, kandidatinį įrašą ir užpuoliko referencinį modelį – ne tam, kad pateiktų instrukciją, o kad grėsmę būtų galima tiksliai analizuoti, o ne aptarti miglotai.

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.

Atakai nereikia specialios privatumo sąsajos. Įprasta diagnostinė užklausa gali išduoti narystės signalą, jei modelis neįprastai užtikrintas dėl anksčiau matyto įrašo.

Original Aurora diagram — The Clean Paper CC BY 4.0

Ką jie nustatė

Trys rezultatai, kiekvienas aštresnis už ankstesnį.

Vidurkiai paslepia labiausiai pažeidžiamus. Vertinant bendrai – kaip įprasta – daugelis modelių atrodo gana privatūs: daugumai pacientų ataka neveikia geriau už atsitiktinį spėjimą. Tačiau kiekvieno paciento lygmens vaizdas kitoks. Kaip Knolle sakė pranešime apie tyrimą, ankstesni vertinimai „visada matavo tik vidutinę riziką visiems pacientams. Mes pirmą kartą ištyrėme riziką atskirų pacientų lygmeniu – ir vaizdas pasirodė visiškai kitoks.“ Kai kuriems žmonėms ataka pavyksta **beveik idealiai** – autorių matuotas atakos AUC yra **0,95 ar didesnis** (0,5 atitinka monetos metimą, 1,0 – nepriekaištingą atskyrimą) – nors viso duomenų rinkinio vidurkis atrodo ne geresnis už atsitiktinumą.



Vidurkis gali būti arti atsitiktinio spėjimo, nors nedidelė įrašų dalis dešiniajame pasiskirstymo „uodegos“ gale išlieka gerokai lengviau atpažįstama. Tyrimo mintis: privatumo riziką reikia tikrinti paciento, o ne vien viso rinkinio lygmeniu.

Original Aurora diagram — The Clean Paper CC BY 4.0

Rizika pasiskirsto nevienodai ir labiausiai tenka nepakankamai atstovaujamoms grupėms. Pacientai, kuriems beveik tobula ataka pavykdavo dažniausiai, sistemingai priklausė duomenyse **nepakankamai atstovaujamoms grupėms**: etninėms mažumoms, retų ligų fenotipams, neįprastų vaizdų grupėms. Elektroninių sveikatos įrašų rinkinyje juodaodžiai pacientai tarp pažeidžiamiausių įrašų pasitaikė **31 % dažniau**, nei būtų tikėtasi; mamografijos rinkinyje vaizdai, pažymėti kaip įtartini dėl piktybiškumo, pavojingiausioje grupėje buvo pertekliniai net **+1 179 %**. (Šie du skaičiai yra pavyzdžiai, o ne visas rezultatas – panašus poslinkis rastas ir kitose grupėse. Be to, tai *santykinis*

perteklius mažoje kraštutinės rizikos grupėje: jis nusako, kiek dažniau šie pacientai pasirodo tarp labiausiai pažeidžiamų, o ne kokia dalis visų juodaodžių ar įtartinų mamogramų pacientų yra atskleidžiama.) Kai duomenyse yra mažiau panašių pavyzdžių, konkretaus žmogaus įrašas labiau išsiskiria – o būtent išskirtinumą tokia ataka ir išnaudoja. Privatumo nesėkmė nėra atsitiktinė; ji telkiasi tarp žmonių, kurie jau yra pakraščiuose.

Didesni modeliai problemą stiprina. Pacientų, kuriems pavyksta beveik tobula ataka, skaičius ryškiai augo didėjant modelio talpai. Viename dermatologijos duomenų rinkinyje ši dalis pakilo nuo beveik nulio mažiausiame modelyje iki maždaug **vieno iš dešimties** didžiausiame. Kadangi medicininiai DI modeliai vis didėja ir tampa pajėgesni – būtent tokia kryptimi juda visa sritis – autoriai perspėja, kad ši konkreti rizika gali stiprėti, o ne nykti.

Rückert išvadą suformulavo tiesiai: „Tai nėra toleruotina rizika. Sveikatos duomenys yra itin jautrūs.“

Kodėl „vidutiniškai saugu“ yra netinkamas testas

Mažas vidutinis rizikos rodiklis lengvai suvokiamas kaip švarus sveikatos pažymėjimas. Pagrindinė darbo pamoka – šiuo atveju vidurkis nėra vien netikslus; jis matuoja ne tą dalyką.

Privatumo garantija prasminga tik tada, jei galioja ir didžiausios rizikos žmogui, o ne vien „vidutiniam“ pacientui. Modelio, kuriame 999 iš 1000 žmonių beveik neįmanoma atpažinti, bet vieną galima nustatyti beveik be klaidos, negalima prasmingai vadinti „99,9 % privačiu“ tam vienam žmogui. O jei tas vienas nuspėjamai yra retos ligos pacientas ar etninės mažumos atstovas, rodiklis ne tik neišsamus – jis tyliai diskriminuoja. Saugumas daugumai paskelbiamas saugumu visiems.

Būtent tokį klausimo poslinkį verčia atlikti šis tyrimas: nuo *kiek privatus modelis vidutiniškai?* prie *kuris pacientas labiausiai pažeidžiamas ir kas jis?* Tai skirtingi klausimai, ir tik antrasis iš tiesų kalba apie individualų privatumą.

Ko tai neįrodo ir neteigia

- Tyrimas **nesako**, kad medicininio DI reikėtų atsisakyti. Autoriai siūlo mažinti riziką, o ne trauktis: ją matuoti ir taisyti, o ne nustoti kurti sistemas.
- Jis **nereiškia**, kad kiekvienas medicininis modelis nutekina informaciją ar kad konkretus įdiegtas modelis jau buvo užpultas. Parodoma, kad *rizika egzistuoja, pasiskirsto nevienodai*, o įprasti bendrieji rodikliai jos nepastebi.
- Jis **nereiškia**, kad jūsų medicinos įrašai jau atskleisti. Atakai reikia konkrečių sąlygų – prieigos prie modelio, kandidatinio įrašo ir užpuoliko referencinės infrastruktūros.
- Jis **nerodo**, kad nuteka paciento įrašo *turinys*. Nuteka narystė – pats įtraukimo į mokymo rinkinį faktas, kuris pavojingas dėl kitų priežasčių.
- Jis **nesusiaurina** nelygybės iki vieno sensacingo skaičiaus. Tvirtas ir pakartojamas rezultatas yra pats *dėsningumas*: bendri rodikliai nepakankamai įvertina individualią riziką, o didžiausia ji tenka

prasčiausiai atstovaujamiems pacientams – ir tai matyti skirtinguose rinkiniuose bei šimtuose modelių.

Kiek tvirti yra įrodymai?

Tai empirinis metodologinis rezultatas, ir jis gana tvirtas. Dėsningumas – kad bendrieji privatumo rodikliai sistemingai nuvertina individualią riziką, likutinė rizika telkiasi nepakankamai atstovaujamose grupėse ir didėja augant modelio talpai – pasikartojo **septyniuose skirtingo tipo duomenų rinkiniuose** ir daugelyje modelių kiekvienam rinkiniui. Būtent tokia apimtis leidžia matavimo išvadą laikyti ne vieno rinkinio artefaktu.

Yra dvi svarbios išlygos. Pirma, tai *rizikos* demonstravimas naudojant pačių autorių sukurtas atakas. Tyrimas parodo, ką gebėtų pakankamai pajėgus užpuolikas esant nurodytoms prielaidoms, bet nematuoja, kaip dažnai tokios atakos vyksta realiame pasaulyje. Antra, įspūdingi skaičiai – beveik tobulas kai kurių pacientų nustatymas – sąmoningai apibūdina pačius blogiausius atvejus. Tai ir yra tyrimo esmė; juos reikia skaityti kaip „pasiskirstymo uodega gerokai sunkesnė, nei rodo vidurkis“, o ne kaip „dauguma pacientų yra atskleidžiami“.

Tinkama reakcija čia nėra nei panika, nei numojimas ranka. Tai kruopštus kelių duomenų rinkinių tyrimas, parodantis, kad standartinis medicininio DI privatumo vertinimas nemato savo blogiausių atvejų – ir kad tie blogiausi atvejai tenka pacientams, kurie turi mažiausiai apsaugos atsargos.

Kodėl tai svarbu

Šį darbą nuo techninės pastabos skiria du dalykai.

Pirma, jis keičia tai, ką turėtų reikšti žodžiai „saugantis privatumą“. Jei modelis išleidžiamas su mažu vidutiniu privatumo rizikos rodikliu, tas vidurkis gali būti visiškai tikras ir vis tiek slėpti pacientų pogrupį, kurio narystę ataka nustato beveik be klaidos. Praktinis autorių reikalavimas konkretus: prieš diegiant modelį vertinti privatumo riziką **kiekvienam pacientui**, kontroliuoti, kas gali naudotis įdiegtu modeliu, ir taikyti tokius metodus kaip **diferencinis privatumas** – mokymo metu įvedamas nedidelis, kruopščiai sukalibruotas triukšmas, silpninantis narystės atakas, už realią ir valdomą modelio naudingumo kainą. Privatumo ir naudingumo kompromisas egzistuoja; jis nėra nemokamas. „Patikrinome vidurkį“ neturėtų reikšti „patikrinome privatumą“.

Antra, tyrimas susieja privatumą su teisingumu. Tos pačios grupės, kurios medicinos duomenyse nepakankamai atstovaujamos ir dėl to jau gali būti prasčiau aptarnaujamos medicininio DI, pasirodo esančios ir prasčiausiai jo privatumo apsaugotos. Sritis, rimtai sprendžianti pirmąją nelygybę, negali antrąją laikyti svetima problema. Tai tie patys žmonės.

Raminanti medicininio DI privatumo istorija – *mes tai išmatavome, vidurkis mažas, vadinasi viskas gerai* – yra būtent ta istorija, kurią šis darbas išardo. Ne tam, kad atbaidytų nuo technologijos, o kad perkeltų standartą ten, kur jam vieta: privatumas yra pažadas, kurį galima teigti ištesėjus tik

patikrinus ir tą žmogų, kuriam žala tikėtiniausia.

Trumpa santrauka

Narystės nustatymo atakos bando išsiaiškinti, ar konkretaus žmogaus įrašas buvo DI modelio mokymo duomenyse. Kadangi pati narystė gali atskleisti jautrią informaciją – pavyzdžiui, kad žmogus sirgo tam tikra liga – tai yra tikras privatumo nutekėjimas net tada, kai pats medicinos įrašas niekada nepasirodo. Ankstesniuose darbuose dažniausiai matuota, kaip dažnai tokia ataka pavyksta *vidutiniškai* visame duomenų rinkinyje. Šiame tyrime rizika matuota *kiekvienam pacientui* septyniuose medicinos duomenų rinkiniuose (vaizdai, EKG, elektroniniai įrašai) ir daugelyje modelių. Rezultatas: bendrieji rodikliai smarkiai nuvertina riziką – kai kuriuos žmones galima nustatyti beveik idealiai (atakos AUC $\geq$ 0,95), nors viso rinkinio vidurkis atrodo saugus; labiausiai pažeidžiami sistemingai yra nepakankamai atstovaujamų grupių pacientai (etninės mažumos, retos ligos, neįprasti vaizdiniai požymiai); o problema didėja augant modelio dydžiui. Autoriai nesiūlo atsisakyti medicininio DI – jie siūlo matuoti privatumą individualiu lygmeniu, kontroliuoti prieigą prie modelių ir taikyti diferencinį privatumą. Pagrindinė išvada: „vidutiniškai privatus“ nėra privatumo garantija, o labiausiai ji nesuveikia tiems, kurie jau ir taip apsaugoti prasčiausiai.

Be pagražinimų

Ką rodo tyrimas: septyniuose medicinos duomenų rinkiniuose ir daugelyje modelių kiekvienam rinkiniui kai kurių pacientų narystės nustatymo rizika yra gerokai didesnė, nei rodo bendrieji rodikliai – iki beveik tobulo atakos rezultato (AUC $\geq$ 0,95). Likutinė rizika neproporcingai tenka nepakankamai atstovaujamoms grupėms ir didėja augant modelio talpai.

Kas tikėtina, bet neįrodyta: kad realūs užpuolikai jau naudoja šį metodą prieš klinikoje įdiegtus modelius. Tyrimas parodo *galimybę ir jos pasiskirstymą* esant nurodytoms prielaidoms, bet nematuoja realių incidentų dažnio.

Ko tyrimas nerodo: kad medicininio DI reikia atsisakyti; kad kiekvienas modelis nutekina informaciją ar konkretus įdiegtas modelis jau buvo pažeistas; kad atskleidžiamas paciento įrašo *turinys*, o ne narystė; kad nelygybę galima išreikšti vienu universaliu skaičiumi – tvirtas rezultatas yra pasikartojantis dėsningumas.

Pagrindiniai ribotumai: matuojama blogiausio atvejo rizika naudojant pačių autorių sukurtas atakas, o ne stebimi realūs incidentai; dramatiški skaičiai sąmoningai apibūdina pačius labiausiai pažeidžiamus žmones; konkrečių grupių skirtumai pateikiami kaip nuoseklus kelių duomenų rinkinių modelis, o ne suvedami į vieną skaičių.

Kiek pasitikėti turėtų bendrasis skaitytojas? Tvirtai – kad bendrieji privatumo rodikliai nepakankamai įvertina individualią riziką, kad likutinė rizika telkiasi tarp nepakankamai atstovaujamų pacientų ir kad ji didėja su modelio dydžiu; tai parodyta daugelyje rinkinių ir modelių. Vidutiniškai

– dėl tokių atakų dažnio realiame pasaulyje, kurio šis tyrimas nematuoja. Saugi išvada nėra nei „medicininis DI nutekina jūsų duomenis“, nei „privatumas jau išspręstas“. Ji tokia: standartinis privatumo patikrinimas nemato savo blogiausių atvejų, o tie atvejai tenka pažeidžiamiausiems pacientams – todėl reikia keisti patį patikrinimą.

Šaltiniai

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