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

# Meditsiiniline AI võib olla keskmiselt privaatne ja siiski paljastada konkreetseid patsiente — eriti alaesindatuid

*„Privaatsust säilitava" meditsiinilise AI mudeli väide toetub tavaliselt ühele keskmisele arvule. See uuring väidab, et see on vale test. Autorid uurisid treeningandmestikku kuulumise järeldamise ründeid — mis paljastavad, kas konkreetse inimese kirje oli mudeli treeningandmetes ja võivad seeläbi reeta näiteks tema haiguse — ning mõõtsid riski koondkeskmise asemel patsiendi kaupa, seitsmes meditsiiniandmestikus (pildid, EKG, elektroonilised terviseandmed) ja paljudes mudelites. Muster: keskmiselt turvalisena näiv mudel võib lasta ründajal konkreetse inimese peaaegu täiuslikult tuvastada (attack AUC $\geq$ 0,95); paljastatud kuuluvad süstemaatiliselt alaesindatud rühmadesse (rahvusvähemus, haruldane haigus, ebatavaline pildistamine); ning probleem kasvab mudelite suurenedes. Autorid ei ütle, et meditsiinilisest AI-st tuleb loobuda — nad ütlevad, et privaatsust tuleb mõõta patsiendi kaupa, mudelile ligipääsu piirata ja kasutada diferentsiaalprivaatsust. Ebamugav tuum: „keskmiselt privaatne" ei ole privaatsusgarantii ning ebaõnnestub patsientide puhul, kes on juba kõige vähem kaitstud.*

---

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

# Privaatsusgarantii on lubadus inimesele, mitte keskmisele

Kui meditsiinilise AI mudelit nimetatakse „privaatsust säilitavaks", toetub see väide tavaliselt ühele arvule: kõigi patsientide seas, kelle andmetega mudelit treeniti, on keskmine tõenäosus järeldada mõne konkreetse inimese kuulumist treeningandmestikku väike. See kõlab rahustavalt. Selle töö vaikne ja ebamugav mõte on, et see on vale arv.

Privaatsus ei ole keskmine. See on igale inimesele antud lubadus — et tema kuulumine treeningandmestikku ei tule hiljem tagasi teda paljastama. Ja keskmine võib seda lubadust peaaegu kõigi jaoks pidada, kuid mõne puhul täielikult murda. Teadlased mõõtsid privaatsusriski patsiendi kaupa, lahutusvõimega, mida varem polnud kasutatud, ning leidsid täpselt selle: koondvaates turvalisena näivad mudelid võivad konkreetsete inimeste kuulumise treeningandmestikku peaaegu täiuslikult reeta — ning paljastatavad inimesed on ebaproportsionaalselt sageli need, kelle kaitse on juba niigi kõige nõrgem.

# Mida autorid tegid

Meeskond — eesotsas Moritz Knollega, koos Daniel Rückertiga (mõlemad Müncheni Tehnikaülikoolist) ja Georg Kaissisega (Hasso Plattneri Instituut) ning kolleegidega Imperial College Londonist — võttis tuntud privaatsusohu nimega **membership inference attack** ehk treeningandmestikku kuulumise järeldamise rünne ja muutis küsimust. Selle asemel et küsida „kui sageli õnnestub rünne keskmiselt üle kogu andmestiku?", küsisid nad „kui paljastatud on *see konkreetne patsient*?".

Nad tegid patsiendipõhise analüüsi **seitsmel väljakujunenud meditsiiniandmestikul**, mis hõlmasid väga erinevaid andmetüüpe — meditsiinipildid, elektrokardiogrammid ja elektroonilised terviseandmed — ning iga andmestiku puhul 200 mudelit. Iga patsiendi ja iga mudeli puhul hinnati, kui kindlalt suudaks ründaja otsustada, kas selle inimese kirje oli treeningandmete hulgas, ning seejärel jagati tulemused rühmade kaupa: haigusseisund, enda teatatud rass, kindlustus, sugu ja pildistamisprotokoll.

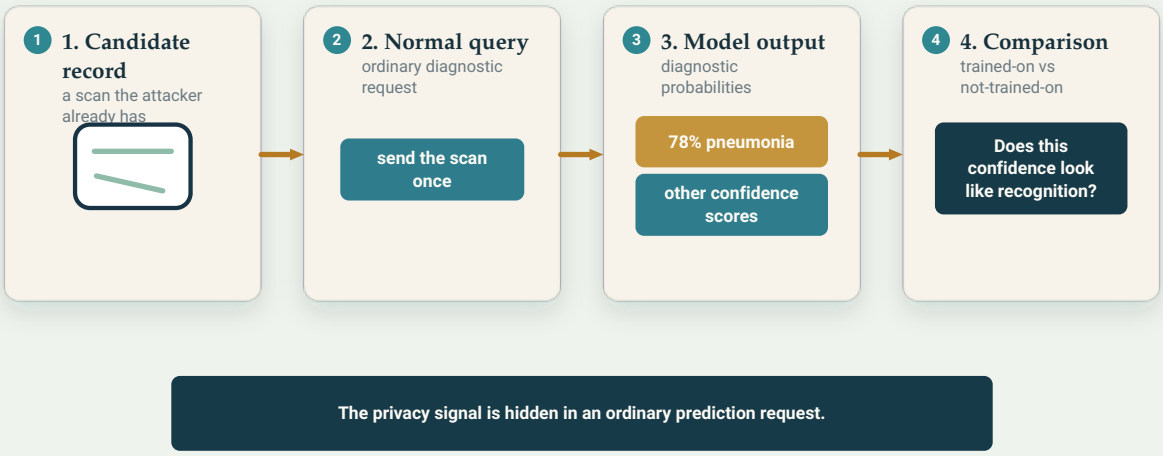
### Mis treeningandmestikku kuulumise järeldamise rünne tegelikult on

Siinne leke on peenem kui „mudel sülitab su haigusloo välja". Jutt käib *kuulumisest* — pelgast faktist, et sinu andmed olid treeningkomplektis.
Alustame sellest, mida selline mudel tavaliselt teeb. Annad talle pildi — näiteks rindkere röntgeni — ja mudel annab tagasi tõenäosused: 78% pneumoonia tõenäosust, lisaks hinnangud kardiomegaalia, turse ja konsolidatsiooni kohta. Rünne kasutab *ainult* seda igapäevast diagnostilist vastust. Ei midagi eksootilist.
Nõrkus, millele rünne toetub, on see: mudel on tavaliselt veidi **enesekindlam** täpselt nende näidete puhul, millel teda treeniti, kui nende puhul, mida ta pole kunagi näinud. Kujuta ette õpilast, kes nägi eksamit salaja ette — harjutatud küsimustele tulevad vastused veidi liiga kiiresti ja veidi liiga enesekindlalt. Selle vihje põhjal kindlakstegemine, kas konkreetne kirje

oli üks mudeli õpitud näidetest, ongi „membership inference".
Rünne käib samm-sammult nii. Keegi tahab teada, kas konkreetse inimese pilti kasutati mudeli treenimiseks. Nad **ei** küsi mudelilt haiguslugu — neil on kandidaadiks olev pilt juba olemas, inimese enda oma või väga lähedane koopia. Nad saadavad selle ühe korra mudelisse tavalise diagnostilise päringuna ja märgivad, kui kindel mudel vastuses on. Siis kontrollivad nad, kas enesekindlus meenutab rohkem mudelit, mis *oli* pildil treenitud, või mudelit, mis *polnud* — seda võrdlust saab odavalt teha, treenides tavalisel arvutil ilma eririistvarata enda asendusmudeli ehk „referentsmudeli", mis õpib ära sellise liigse enesekindluse tunnuse. Kui pärismudel on selle pildi puhul ebatavaliselt kindel — justkui tunneks selle ära — on see tugev tõend, et pilt oli treeningandmetes.
Häiriv osa on see, et päring ei näe üldse ründe moodi välja: see on sama diagnostiline küsimus, mida esitaks klinitsist, ja privaatsussignaal on peidus tavalise ennustuse sees. Just seepärast on risk konkreetne — kuigi seni on seda demonstreeritud kirjeldatud laborieelduste all, mitte päris ründes looduses.
Miks kuulumine üldse loeb, kui haiguslugu ise ei leki? Sest *kuulumine on fakt*. Kui mudelit treeniti patsientidel, kes said kindlat vähi immunoteraapiat, siis kinnitamine, et sinu kirje on mudeli treeningandmetes, reedab, et sul tõenäoliselt oli see vähk — asi, mida kindlustaja või tööandja ei tohiks kunagi järeldada saada. Sisu jääb suletuks; välja pääseb kuulumise fakt.
Autorid panevad need eeldused selgelt kirja — ligipääs mudeli tavalistele ennustustele, kandidaadikirje ja ründaja enda referentsmudel — mitte õpetusena, vaid selleks, et ohtu saaks põhjendatult analüüsida, mitte käega kõrvale lükata.

## 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.*

Rünne ei vaja erilist privaatsus-API-t. Tavaline diagnostiline päring võib lekkida kuulumissignaali, kui mudel on varem nähtud kirje suhtes ebatavaliselt enesekindel.

# Mida nad leidsid

Kolm tulemust, iga järgmine teravam kui eelmine.

**Keskmised peidavad paljastatud inimesed.** Koondtasandil — tavapärasel mõõtmisviisil — näivad paljud mudelid rahustavalt privaatsed: enamiku patsientide puhul ei saa rünne juhusest paremat tulemust. Patsiendipõhine vaade rääkis teise loo. Nagu Knolle sõnastas uuringu teadaandes, olid varasemad hinnangud „mõõtnud ainult keskmist riski kõigi patsientide seas. Me uurisime esimest korda riski üksikpatsiendi tasandil — ja pilt on väga erinev." Mõne inimese puhul õnnestub rünne **peaaegu täiuslikult** — autorid mõõdavad attack AUC väärtuseks **0,95 või rohkem** (0,5 on mündivise, 1,0 täiuslik) — isegi siis, kui kogu andmestiku keskmine ei paista juhusest parem.


Keskmine võib olla juhuslikule tasemele lähedal, samal ajal kui väike kirjete „saba" jääb palju rohkem paljastatuks. Töö mõte on, et privaatsust tuleb kontrollida patsiendi tasandil, mitte ainult koondnäitajana.



**Paljastatus on ebavõrdne — ja langeb alaesindatud rühmadele.** Peaaegu täiusliku ründe suhtes kõige haavatavamad patsiendid kuulusid süstemaatiliselt andmetes **alaesindatud rühmadesse**: vähemusrahvused, haruldaste haiguste fenotüübid, ebatavalised pildistamistunnused. Elektrooniliste terviseandmete andmestikus esinesid mustanahalised patsiendid kõige haavatavamate kirjete seas **31% sagedamini** kui oodata; mammograafiaandmestikus olid pahaloomulisuse suhtes kahtlaseks märgitud uuringud selles riskitsoonis üleesindatud lausa **+1179%**. (Need kaks arvu on näited, mitte kogu

pilt — töö näitab sama kallutatust mitmes rühmas — ning need on *suhtelised* üleesindatused väikeses äärmusliku riskiga sabas: kui palju sagedamini neid patsiente kõige paljastatumate kirjete seas leidub, mitte mustanahaliste või kahtlase mammograafiaga patsientide osakaal, kes on paljastatud.) Mudelil on vähem sarnaseid näiteid, mille sisse need patsiendid ära hajuksid, mistõttu nende kirjed paistavad välja — ja just silmapaistvust kuulumise järeldamise rünne otsib. Privaatsuse läbikukkumine pole juhuslik; see koondub inimestele, kes on juba äärealal.

**Suuremad mudelid teevad asja hullemaks.** Peaaegu täiuslike rünnete suhtes paljastatud patsientide arv kasvas mudeli võimekusega järsult — ühes dermatoloogiaandmestikus tõusis osakaal väikseima mudeli sisuliselt nullist suurimas mudelis umbes **üheni kümnest**. Kuna meditsiinilise AI mudelid muutuvad suuremaks ja võimekamaks — just selles suunas liigub kogu valdkond — muutub see konkreetne risk autorite hoiatuse järgi tõsisemaks, mitte väiksemaks.

Rückerti kokkuvõte on järsk: „See ei ole talutav risk. Terviseandmed on väga tundlikud."

## Miks „keskmiselt turvaline" on vale test

Kiusatus on lugeda väike keskmine risk puhta tervisetõendina. Töö keskne õppetund on, et keskmistamine pole siin lihtsalt ebatäpne — see mõõdab valet asja.

Privaatsusgarantii on tähenduslik ainult siis, kui see kehtib kõige suurema riskiga inimese, mitte keskmise inimese jaoks. Mudel, kus 999 patsienti 1000-st pole tuvastatavad, kuid ühe saab peaaegu täiuslikult kindlaks teha, pole selle ühe inimese jaoks mingis tähenduslikus mõttes „99,9% privaatne" — ning kui see üks inimene on ennustatavalt haruldase haiguse või rahvusvähemuse patsient, pole mõõdik mitte ainult puudulik, vaid vaikselt diskrimineeriv. See teatab enamuse turvalisusest ja nimetab seda kõigi turvalisuseks.

Just selle nihke töö sunnib tegema: küsimuselt *kui privaatne on see mudel keskmiselt?* küsimusele *kes on kõige paljastatum patsient ja kes ta on?* Need on erinevad küsimused ning ainult teine on päriselt privaatsusküsimus.

## Mida see *ei* tõesta — ega väida

- See **ei** ütle, et meditsiinilisest AI-st tuleks loobuda. Autorite raamistik on riski vähendamine, mitte taganemine: mõõda ja paranda risk, ära lõpeta ehitamist.
- See **ei** tähenda, et iga meditsiinimudel lekib või et mõnd konkreetset kasutuses olevat mudelit on rünnatud. See näitab, et *risk on olemas ja jaotub ebavõrdselt* ning standardsed koondmõõdikud ei näe seda.
- See **ei** tähenda, et sinu andmed on juba paljastatud. Rünne vajab konkreetseid tingimusi — ligipääsu mudelile, kandidaadikirjet ja ründaja enda taristut — mitte juhuslikku võimekust.
- See **ei** näita, et patsiendi *sisu* lekib. Lekib kuulumine — kaasamise fakt — mis on ohtlik teisel põhjusel, mitte sellepärast, et haiguslugu välja prinditakse.

- See **ei** taanda ebavõrdsust ühele löövale arvule. Tugev ja korratav tulemus on *muster* — koondmõõdikud alahindavad individuaalset riski ning alaesindatud patsiendid kannavad sellest suurimat osa — üle andmestike ja sadade mudelite.

## Kui tugevad on tõendid?

See on empiiriline metodoloogiline tulemus ja robustne. Muster — koondprivaatsusmõõdikud alahindavad süstemaatiliselt patsiendipõhist riski, jääkrisk koondub alaesindatud rühmadele ja kasvab mudeli võimekusega — püsis **seitsmes eri tüüpi andmetega andmestikus** ja suure arvu mudelite puhul igas andmestikus. Just selline laius teeb mõõtmisväite usutavaks, mitte ühe andmestiku artefaktiks.

Kaks ausat reservatsiooni. Esiteks on see *riski* demonstratsioon, mida mõõdetakse autorite endi ehitatud rünnetega; see kirjeldab, kui hästi võimekas ründaja *võiks* antud eeldustel hakkama saada, mitte seda, kui sageli päris ründeid toimub. Teiseks kirjeldavad silmatorkavad arvud — mõne patsiendi peaaegu täiuslik tuvastamine — *disaini järgi* kõige halvemas olukorras inimesi; see ongi kogu mõte ning neid tuleb lugeda kui „saba on palju raskem, kui keskmine lubab arvata", mitte kui „enamik patsiente on paljastatud".

Sobiv hoiak pole ei paanika ega mahategemine: see on hoolikas mitme andmestikuga uuring, mis näitab, et standardne viis meditsiinilise AI privaatsust sertifitseerida on pime omaenda halvimate juhtumite suhtes — ning need juhtumid langevad patsientidele, kellel on kõige vähem puhvrit.

## Miks see oluline on

Kaks asja teevad sellest rohkem kui tehnilise joonealuse märkuse.

Esiteks muudab see seda, mida „privaatsust säilitaval" peaks üldse lubama tähendada. Kui mudel avaldatakse keskmise privaatsusskooriga, võib see skoor olla päriselt väike ja ikkagi peita patsientide alamhulga, kes on kuulumise järeldamise ründega peaaegu täiuslikult tuvastatavad. Töö praktiline nõue on konkreetne: hinnake privaatsusriski enne avaldamist **patsiendi kaupa**, kontrollige, kellel on ligipääs kasutuses olevatele mudelitele, ja kasutage tehnikaid nagu **diferentsiaalprivaatsus** — väike, hoolikalt kalibreeritud müra, mida lisatakse treeningul ja mis nõrgestab kuulumisrünnet, reaalse ja juhitava hinnaga mudeli kasulikkusele (privaatsuse ja kasulikkuse kompromiss on selge, mitte tasuta). „Me kontrollisime keskmist" ei tohiks enam tähendada, et kontroll tehti.

Teiseks põimib see privaatsuse õiglusega. Samad rühmad, kes on meditsiiniandmetes alaesindatud — ja keda meditsiiniline AI seetõttu juba halvemini teenindab — osutuvad nendeks, kelle privaatsust sama AI kõige vähem kaitseb. Valdkond, mis töötab kõvasti esimese ebavõrdsuse kallal, ei saa käsitleda teist kellegi teise osakonnana. Need on samad inimesed.

Meditsiinilise AI privaatsuse rahustava loo — *me mõõtsime, keskmine on madal, järelikult on kõik korras* — võtab see töö lahti. Mitte selleks, et kedagi tehnoloogiast eemale hirmutada, vaid et tõsta standard

sinna, kuhu see kuulub: lubadus, mille pidamist saab väita alles pärast seda, kui oled seda kontrollinud inimese jaoks, kes võib kõige rohkem haiget saada.

# Lühidalt

Treeningandmestikku kuulumise järeldamise ründed püüavad kindlaks teha, kas konkreetse inimese kirje oli AI mudeli treeningandmetes — ja kuna kuulumine võib ise olla paljastav, näiteks reeta, et inimesel oli kindel haigus, on see päris privaatsusleke isegi siis, kui haigusloo sisu kunagi ei ilmu. Varasem töö mõõtis selliste rünnete edukust *keskmiselt* üle andmestiku. See uuring mõõtis seda *patsiendi kaupa* seitsmes meditsiiniandmestikus (pildid, EKG, elektroonilised terviseandmed) ja paljudes mudelites ning leidis, et koondmõõdikud alahindavad riski tugevalt: mõne inimese saab peaaegu täiuslikult tuvastada (attack AUC $\geq$ 0,95) isegi siis, kui kogu andmestiku keskmine näib turvaline; kõige haavatavamad kuuluvad süstemaatiliselt alaesindatud rühmadesse (rahvusvähemus, haruldane haigus, ebatavaline pildistamine); ning probleem kasvab mudeli suurusega. Autorid ei väida, et meditsiinilisest AI-st tuleks loobuda — nad soovitavad mõõta privaatsust individuaalsel tasandil, piirata mudelile ligipääsu ja kasutada diferentsiaalprivaatsust. Järeldus: „keskmiselt privaatne" ei ole privaatsusgarantii ning see ebaõnnestub enamasti just patsientide puhul, kes on juba kõige vähem kaitstud.

# Ilustamata kontroll

**Mida töö näitab:** Seitsmes meditsiiniandmestikus ja paljudes mudelites on patsiendipõhine treeningandmestikku kuulumise järeldamise risk mõne inimese jaoks palju suurem, kui koondmõõdikud näitavad — kuni peaaegu täiusliku ründeeduni (AUC $\geq$ 0,95) — ning jääkrisk langeb ebaproportsionaalselt alaesindatud rühmadele ja kasvab mudeli võimekusega.

**Mis on usutav, kuid tõestamata:** Et päris ründajad kasutavad seda juba kasutuses olevate kliiniliste mudelite vastu; uuring demonstreerib *võimekust ja selle jaotust* antud eeldustel, mitte päris rünnete sagedust.

**Mida see ei näita:** Et meditsiinilisest AI-st tuleks loobuda; et iga mudel lekib või mõnd konkreetset kasutuses olevat mudelit on murtud; et paljastub patsiendi haigusloo *sisu* ja mitte ainult kuulumine; ühte kvantifitseeritud ebavõrdsusarvu — robustne tulemus on muster, mitte üks number.

**Peamised piirangud:** Mõõdetakse halvima juhu riski autorite endi rünnete kaudu, mitte vaadeldud intsidente; dramaatilised arvud kirjeldavad disaini järgi kõige paljastatumaid inimesi; ning rühmadevahelisi täpseid erinevusi näidatakse järjekindla mustrina üle andmestike, mitte ei taandata ühele arvule.

**Kui palju võiks tavalugeja seda usaldada?** Kõrge kindlus, et koondprivaatsusmõõdikud alahindavad individuaalset riski, et jääkrisk koondub alaesindatud patsientidele ja et see halveneb mudeli suurusega — seda on näidatud paljude andmestike ja mudelite peal. Mõõdukas kindlus selliste rünnete

päriselu sageduse suhtes, mida töö ei mõõda. Turvaline lugemine pole „meditsiiniline AI lekitab su andmeid" ega „privaatsus on lahendatud", vaid „standardne privaatsuskontroll ei näe omaenda halvimaid juhtumeid ning need langevad kõige haavatavamatele patsientidele — seega peab kontroll muutuma".

## Allikad

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