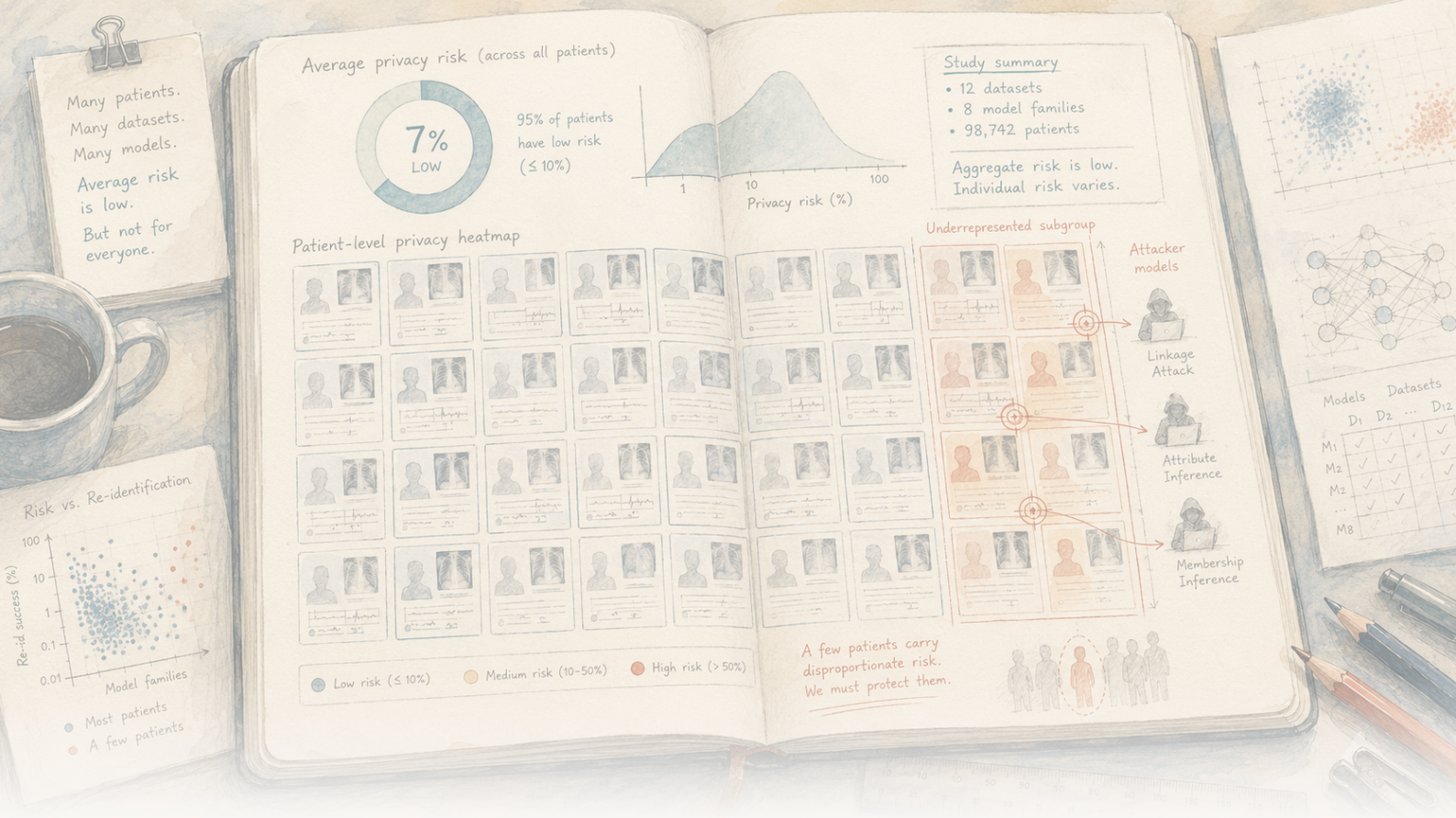


The Clean Paper

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

Persónuverndartrygging er loforð við einstakling, ekki meðaltal

Rannsóknin mælir persónuverndaráhættu sjúklinga hvern fyrir sig í stað þess að treysta einu meðaltali fyrir líkanið. Í sjö læknisfræðilegum gagnasöfnum gátu líkön sem virtust örugg að meðaltali samt gert suma einstaklinga næstum fullkomlega auðkennanlega í membership-inference árás, sérstaklega fólk úr vanfulltrúuðum hópum. „Persónuvernd að meðaltali“ er því ekki trygging fyrir hvern sjúkling.

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

Persónuverndartrygging er loforð við einstakling, ekki meðaltal

Þegar sagt er að læknisfræðilegt gervigreindarlíkan „verndi persónuvernd“ byggir sú fullyrðing oft á einni tölu: að meðaltali sé lítil hætta á að hægt sé að ályktа hvort tiltekinn einstaklingur hafi verið í þjálfunargögnunum. Það hljómar traustvekjandi. Rólega og óþægilega niðurstaða þessarar greinar er að þetta sé **ranga talan**.

Persónuvernd er ekki meðaltal. Hún er loforð til hvers einstaklings um að þátttaka hans í þjálfunargögnum leiði ekki síðar til þess að hægt sé að afhjúpa upplýsingar um hann. Meðaltal getur staðið við það loforð fyrir næstum alla og samt brugðist algjörlega gagnvart fáeinum. Höfundarnir mældu áhættuna **einn sjúkling í einu**, með nákvæmni sem áður hafði ekki verið notuð, og fundu einmitt þetta: líkön sem virðast örugg í heild geta afhjúpað aðild ákveðinna einstaklinga næstum fullkomlega — og þeir einstaklingar eru óhóflega oft úr hópum sem þegar njóta minnstu verndar.

Hvað gerðu höfundarnir?

Teymið, undir forystu Moritz Knolle ásamt Daniel Rückert við Technical University of Munich, Georg Kaissis við Hasso Plattner Institute og samstarfsfólki við Imperial College London, tók vel þekkta persónuverndarógn sem kallast **aðildarályktunarárás** (*membership inference attack*) — árás sem reynir að ráða af niðurstöðum líkans hvort tiltekin skrá hafi verið í þjálfunarsafninu — og breytti spurningunni. Í stað „hversu oft tekst árásin að meðaltali yfir allt gagnasafnið?“ spurðu þau: „hversu berskjaldaður er **þessi tiltekni sjúklingur**?“

Greiningin var gerð á **sjö rótgrónum læknisfræðilegum gagnasöfnum** sem ná yfir mjög ólík gögn — myndgreiningu, hjartarafrit og rafrænar sjúkraskrár — og yfir fjölda líkana, með allt að 200 líkönum á gagnasafn. Fyrir hvern sjúkling í hverju líkani mátu höfundarnir hversu örugglega árásaraðili gæti sagt hvort skrá viðkomandi hefði verið í þjálfunargögnunum. Síðan brutu þeir niðurstöðurnar niður eftir hópum: sjúkdómsstöðu, sjálfskráðri kynþátt/etnísku, tryggingastöðu, kyni og myndgreiningarferli.

> **Hvað aðildarályktunarárás er í raun**
>
> Lekinn hér er lúmskari en „líkanið spýtir út sjúkraskránni“. Spurningin er **aðildin sjálf** — sú staðreynd að gögn einstaklings voru í þjálfunarsafninu.
> Byrjum á venjulegri notkun líkans. Þú gefur því mynd, til dæmis röntgenmynd af brjóstkassa, og það skilar líkum: 78% líkur á lungnabólgu ásamt mati á hjartastækkun, bjúg og þéttingu. **Þetta venjulega greiningarsvar er það eina sem árásin notar.** Enginn sérstakur „persónuverndarhamur“ er nauðsynlegur.
> Veikleiki sem árásin nýtir er að líkan er oft aðeins **öruggara á nákvæmlega þeim dæmum sem það var þjálfað á** en dæmum sem það hefur aldrei séð. Hugsaðu þér nemanda sem sá prófið leynilega fyrirfram: á spurningunum sem hann hafði æft koma svörin örlítið hraðar og örlítið

öruggari. Að ályktа af þessu hvort ákveðin skrá hafi verið meðal þjálfunardæmanna er það sem „aðildarályktun“ (*membership inference*) merkir.

Árásin gengur því svona fyrir sig. Einhver vill vita hvort tiltekin mynd einstaklings var notuð til þjálfunar. Hann biður **ekki** líkanið um sjúkraskrána — hann er þegar með umsóknarmyndina. Hann sendir hana inn sem venjulega greiningarbeiðni og skráir hversu öruggt svarið er. Síðan ber hann þá öryggistilfinningu saman við hegðun líkana sem höfðu eða höfðu ekki séð svipað dæmi. Þetta er hægt að læra með einföldu eigin samanburðarlíkani. Ef hið raunverulega líkan er óvenju öruggt á þessari mynd — eins og það „þekki“ hana — er það vísbending um að myndin hafi verið í þjálfunarsafninu.

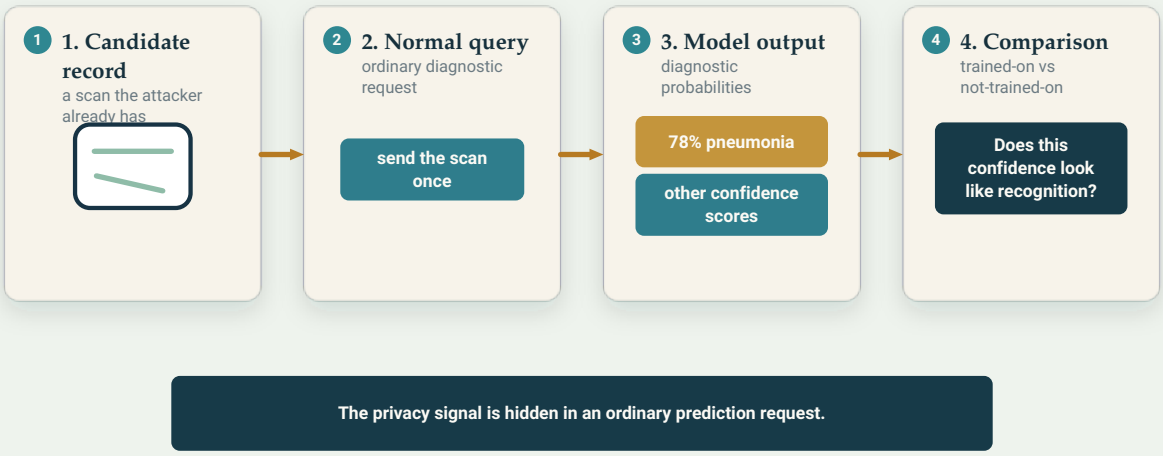
Óþægilegi hlutinn er að beiðnin sjálf lítur ekki út eins og árás: hún er sama greiningarbeiðni og heilbrigðisstarfsmaður gæti sent. Persónuverndarmerkið felst í venjulegri spá. Þess vegna er áhættan raunveruleg, þótt hún hafi hér verið sýnd undir skilgreindum rannsóknarstofuforsendum en ekki skráð sem villt árás á raunverulegt kerfi.

Af hverju skiptir aðild máli ef innihald skrárinnar lekur aldrei? Vegna þess að **aðild er sjálf staðreynd**. Ef líkan var þjálfað á sjúklingum sem fengu ákveðna krabbameinsónæmismeðferð gæti staðfesting á því að skrá þín sé í því gefið mjög sterka vísbendingu um að þú hafir haft þann sjúkdóm. Þetta er tegund ályktunar sem tryggingafélag eða vinnuveitandi ætti ekki að geta gert. Innihaldið helst lokað; staðreyndin um að tilheyra hópnum er það sem sleppur út.

Höfundarnir setja forsendurnar fram skýrt — aðgangur að venjulegum spám líkansins, ein tiltekin hugsanleg sjúkraskrá og eigið samanburðarlíkan árásaraðila — ekki sem leiðbeiningar, heldur svo hægt sé að ræða ógnina nákvæmlega.

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.

Árásin þarf ekki sérstakt persónuverndarviðmót. Venjuleg greiningarbeiðni getur borið merki um aðild að þjálfunarsafni ef líkanið er óvenju öruggt á skrá sem það hefur séð áður.

Hvað fundu þeir?

Þrjár niðurstöður, hver skarpari en sú fyrri.

Meðaltöl fela þá sem eru berskjaldaðir. Þegar áhættan er mæld samanlagt — eins og venjan er — virðast mörg líkön mjög örugg: árásin er ekki betri en tilviljun fyrir flesta sjúklinga. Greining einstaklings fyrir einstakling sýndi annað. Eins og Knolle orðaði það í tilkynningu rannsóknarinnar, höfðu fyrri úttektir aðeins mælt meðaláhættu yfir alla sjúklinga; þegar áhættan var skoðuð á einstaklingsstigi kom allt önnur mynd fram. Fyrir suma einstaklinga tekst árásin **næstum fullkomlega** — **AUC-gildi árásarinnar 0,95 eða hærra**, þar sem 0,5 er myntkast og 1,0 fullkomin aðgreining — jafnvel þegar meðaltal alls gagnasafnsins lítur út eins og tilviljun.

Meðaltalið getur legið nálægt tilviljun á meðan lítill „hali“ skráa er miklu berskjaldaðri. Kjarni greinarinnar er að persónuvernd verður að mæla á sjúklingastigi, ekki aðeins samanlagt.

Áhættan er ójöfn — og lendir á vanfulltrúuðum hópum. Sjúklingarnir sem voru viðkvæmastir fyrir næstum fullkominni árás tilheyrðu kerfisbundið hópum sem voru **vanfulltrúaðir í gögnunum**: minnihlutahópum, sjaldgæfum sjúkdómsgerðum og óvenjulegum myndgreiningareiginleikum. Í gagnasafni rafrænna sjúkraskráa komu svartir sjúklingar **31% oftar** en búast mætti við meðal mest berskjölduðu skráanna. Í brjóstamyndasafni voru myndir sem merktar voru grunsamlegar um illkynja sjúkdóm yfirfulltrúaðar í hættuhalanum um **+1.179%**. Þessar tölur eru dæmi, ekki

heildarmyndin, og þær lýsa **hlutfallslegri yfirfulltrúa innan litla hópsins með mest áhættu**, ekki því hlutfalli allra svartra sjúklinga eða allra grunsamlegra myndataka sem eru berskjaldaðar.

Líkan hefur færri sambærileg dæmi til að „fela“ slíkan sjúkling meðal. Skráin stendur því meira út — og það er einmitt það sem aðildarályktun nemur. Persónuverndarbresturinn er ekki tilviljunarkenndur; hann safnast á fólk sem þegar er á jaðrinum.

Stærri líkön gera vandann verri. Fjöldi sjúklinga sem voru berskjaldaðir fyrir nær fullkominni árás jókst mjög með getu og stærð líkans. Í einu húðsjúkdómsgagnasafni fór hlutfallið frá nánast núlli í minnsta líkaninu upp í um **einn af hverjum tíu** í stærsta líkaninu. Þegar læknisfræðileg AI-líkön verða stærri og öflugri — sem er stefna sviðsins — verður þessi sérstaka áhætta, samkvæmt höfundunum, meiri en ekki minni.

Samantekt Rückert er beinskeytt: „Þetta er ekki ásættanleg áhætta. Heilbrigðisgögn eru mjög viðkvæm.“

Af hverju „öruggt að meðaltali“ er rangt próf

Freistandi er að lesa lága meðaláhættu sem heilbrigðisvottorð. Kjarni greinarinnar er að meðaltalið sé hér ekki aðeins ónákvæmt — það mælir **ranga hlutinn**.

Persónuverndartrygging hefur aðeins merkingu ef hún gildir fyrir einstaklinginn sem er í mestri hættu, ekki aðeins þann sem situr í miðju dreifingarinnar. Líkan þar sem 999 af 1.000 sjúklingum eru nær ógreinanlegir en einn er auðgreindur með mikilli nákvæmni er ekki „99,9% persónuverndað“ í neinum skilningi sem skiptir þeim eina einstaklingi máli. Og ef sá einstaklingur er fyrirsjáanlega sjúklingur með sjaldgæfan sjúkdóm eða úr minnihlutahópi er mælikvarðinn ekki bara ófullkominn heldur einnig hljóðlega mismunandi milli hópa. Hann tilkynnir öryggi meirihlutans og kallar það öryggi allra.

Það er breytingin sem greinin krefst: frá **„hversu private er líkanið að meðaltali?“** yfir í **„hver er mest berskjaldaði sjúklingurinn, og hvaða hópi tilheyrir hann?“** Þetta eru ólíkar spurningar og aðeins sú seinni er raunverulega einstaklingsbundin persónuverndarspurning.

Hvað sýnir þetta *ekki* — eða heldur ekki fram?

- Það segir **ekki** að hætta eigi þróun læknisfræðilegrar gervigreindar. Höfundarnir tala um að draga úr áhættu, ekki hörfa frá tækninni.
- Það þýðir **ekki** að öll læknislíkön leki eða að ákveðið kerfi í notkun hafi verið ráðist á. Það sýnir að **áhættan er til og dreifist ójafnt**, og að hefðbundin meðalviðmið missa af henni.
- Það þýðir **ekki** að sjúkraskrár séu nú þegar opinberar. Árásin þarf ákveðin skilyrði — aðgang að líkaninu, hugsanlega skrá og eigin innviði árásaraðila.
- Það sýnir **ekki** að innihald sjúkraskrár leki. Það sem lekur er **aðildin**, staðreyndin um að skrá hafi verið í þjálfunargögnunum, sem er hættulegt af annarri ástæðu.

- Það dregur **ekki** alla mismunun niður í eina töluna. Sterka og endurtekanlega niðurstaðan er mynstrið: samanlögð viðmið vanmeta einstaklingsáhættu og vanfulltrúaðir sjúklingar bera stærstan hluta af henni.

Hversu sterk eru gögnin?

Þetta er empírísk og aðferðafræðileg niðurstaða, og nokkuð traust. Mynstrið — að meðalpersónuvernd vanmeti áhættu einstakra sjúklinga, að eftirstandandi áhætta safnist á vanfulltrúaða hópa og að hún aukist með stærð líkans — kom fram yfir **sjö gagnasöfn með ólíkar gagnategundir** og fjölda líkana. Sú breidd gerir mælingarniðurstöðuna trúverðugri en ef hún kæmi úr einu gagnasafni.

Tvö fyrirvara þarf að halda skýrum. Í fyrsta lagi er þetta sýning á **áhættu**, mæld með árásum sem höfundarnir sjálfir framkvæma. Hún lýsir því hversu vel fær árásaraðili **gæti** gert undir tilgreindum forsendum, ekki hversu oft raunverulegar árásir eiga sér stað.

Í öðru lagi lýsa sláandi tölurnar — nær fullkomin árás á suma — **verstu einstaklingunum samkvæmt hönnun greiningarinnar**. Það er einmitt tilgangurinn. Þær ætti að lesa sem „halinn er miklu þyngri en meðaltalið sýnir“, ekki sem „flestir sjúklingar eru berskjaldaðir“.

Rétta afstaðan er því hvorki skelfing né afgreiðsla: vel unnin rannsókn yfir mörg gagnasöfn sem sýnir að hefðbundna leiðin til að votta persónuvernd læknisfræðilegrar gervigreindar sér ekki eigin verstu tilfelli — og að verstu tilfellin lenda á fólki með minnstan varasjóð.

Af hverju skiptir þetta máli?

Tvö atriði gera þetta að meira en tæknilegri neðanmálsgrein.

Í fyrsta lagi breytir greinin því hvað **„persónuverndandi“** ætti að mega þýða. Líkan getur fengið lágt meðalskor og samt falið lítinn hóp sjúklinga sem eru nær fullkomlega auðkennanlegir með aðildarárásum. Hagnýt krafa greinarinnar er skýr: meta persónuvernd **fyrir hvern sjúkling** fyrir útgáfu, stjórna aðgangi að kerfum í notkun og nota aðferðir eins og **differential privacy** — lítið, vandlega kvarðað suð sem bætt er við þjálfun til að gera aðildarárásum erfiðari, með raunverulegum og stjórnanlegum kostnaði í nytsemi líkansins. Persónuvernd–nytsemisviðskiptin eru raunveruleg, ekki ókeypis. „Við skoðuðum meðaltalið“ ætti ekki lengur að teljast fullnægjandi úttekt.

Í öðru lagi tengir greinin persónuvernd við **sanngirni**. Sömu hópar og eru vanfulltrúaðir í læknisfræðilegum gögnum — og fá því oft verri þjónustu frá gervigreind — reynast líka þeir sem tæknin verndar verst með tilliti til persónuverndar. Svið sem vinnur að fyrra ójafnræðinu getur ekki meðhöndlað hitt sem mál einhvers annars. Þetta eru sömu einstaklingarnir.

Traustvekjandi sagan um persónuvernd læknisfræðilegrar AI — *við mældum hana, meðaltalið er lágt, allt er í lagi* — er sagan sem þessi grein tekur í sundur. Ekki til að hræða fólk frá tækninni, heldur til að færa staðalinn þangað sem hann á heima: loforð sem aðeins má segjast halda ef það hefur verið

prófað á þeim sem er líklegastur til að verða fyrir skaða.

Í stuttu máli

Aðildarályktunarárásir reyna að komast að því hvort skrá ákveðins einstaklings hafi verið í þjálfunargögnum AI-líkans. Þar sem aðildin sjálf getur afhjúpað viðkvæma staðreynd — til dæmis að einstaklingur hafi ákveðinn sjúkdóm — er þetta raunverulegur persónuverndarleki jafnvel þótt sjúkraskráin sjálf birtist aldrei. Fyrri rannsóknir mældu árásarárangur **að meðaltali** yfir gagnasafn. Þessi rannsókn mældi hann **fyrir hvern sjúkling** yfir sjö læknisfræðileg gagnasöfn — myndgreiningu, ECG og rafrænar sjúkraskrár — og mörg líkön. Niðurstaðan er að samanlögð viðmið vanmeti áhættuna verulega: suma einstaklinga er hægt að greina nær fullkomlega (AUC-gildi árásar $\geq$ 0,95) jafnvel þegar meðaltal gagnasafnsins lítur öruggt út; mest berskjölduðu einstaklingarnir tilheyra kerfisbundið vanfulltrúuðum hópum, svo sem minnihlutahópum, sjaldgæfum sjúkdómum og óvenjulegum myndgreiningum; og vandinn vex með stærð líkans. Höfundarnir leggja ekki til að hætta læknisfræðilegri AI heldur að mæla persónuvernd einstaklingsbundið, takmarka aðgang og nota meðal annars mismunavernd (*differential privacy*). Meginniðurstaðan: **„öruggt að meðaltali“ er ekki persónuverndartrygging**, og þegar meðaltalið bregst eru það oft þeir sem þegar eru minnst verndaðir sem bera áhættuna.

Raunhæf yfirferð

Hvað sýnir greinin? Yfir sjö læknisfræðileg gagnasöfn og mörg líkön er membership-inference-áhætta sumra sjúklinga miklu meiri en samanlögð viðmið gefa til kynna — allt að nær fullkominni árás (AUC $\geq$ 0,95) — og eftirstandandi áhætta lendir óhóflega á vanfulltrúuðum hópum og eykst með getu/stærð líkans.

Hvað er trúverðugt en ósannað? Að raunverulegir árásaraðilar séu þegar að nýta þetta gegn klínískum kerfum. Rannsóknin sýnir **getuna og dreifingu hennar** við tilgreindar forsendur, ekki tíðni raunverulegra árása.

Hvað sýnir greinin ekki? Að hætta eigi læknisfræðilegri gervigreind; að hvert líkan leki eða ákveðið kerfi hafi verið brotið; að innihald sjúkraskráa leki fremur en aðild; eða eina tölu sem lýsir allri hópamismunun. Sterka niðurstaðan er mynstrið, ekki eitt prósentugildi.

Helstu takmarkanir: Rannsóknin mælir versta áhættu með eigin árásum höfundanna, ekki raunveruleg atvik; dramatísku tölurnar lýsa þeim mest berskjölduðu samkvæmt hönnun; og nákvæm hópamismunun kemur fram sem endurtekið mynstur milli gagnasafna fremur en ein sameinuð tala.

Hversu mikið traust ætti almennur lesandi að leggja á niðurstöðuna? Mikið traust á að samanlögð persónuverndarviðmið vanmeti einstaklingsáhættu, að eftirstandandi áhætta safnist á vanfulltrúaða sjúklinga og að hún aukist með stærri líkönum — þetta birtist yfir mörg gagnasöfn og líkön. Miðlungs traust á hversu algengar slíkar árásir eru í raunheimum, því það var ekki mælt. Rétta niðurstaðan

er ekki „læknis-AI lekur gögnunum þínum“ og ekki „persónuvernd er leyst“, heldur: **hefðbundna persónuverndarprófið sér ekki eigin verstu tilfelli og prófið þarf að breytast.**

Heimildir

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