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

Lääketieteellinen tekoäly voi olla keskimäärin yksityinen ja silti paljastaa tiettyjä potilaita — etenkin aliedustettuja

Kun lääketieteellistä tekoälymallia kutsutaan 'yksityisyyttä säilyttäväksi', väite nojaa yleensä yhteen keskimääräiseen lukuun. Tämä tutkimus väittää, että se on väärä testi. Tutkiessaan membership inference -hyökkäyksiä — joilla päätellään, kuuluiko tietyn henkilön tieto mallin koulutusdataan ja voidaan siten paljastaa esimerkiksi että hänellä oli tietty sairaus — tutkijat mittasivat riskin potilaskohtaisesti eivätkä vain aggregaattina, seitsemässä lääketieteellisessä aineistossa (kuvantaminen, EKG, sähköiset potilastiedot) ja useissa malleissa. Kuvio oli selvä: keskimäärin turvallisilta näyttävät mallit voivat silti tehdä yksittäisistä henkilöistä lähes täydellisesti tunnistettavia (attack AUC ≥ 0.95); altistuneet kuuluvat järjestelmällisesti aliedustettuihin ryhmiin, kuten etnisiin vähemmistöihin, harvinaissairaisiin ja poikkeavan kuvantamisen potilaisiin; ja ongelma pahenee mallien kasvaessa. Tutkijat eivät sano, että lääketieteellinen tekoäly pitäisi hylätä — he sanovat, että yksityisyyttä pitää mitata potilaskohtaisesti, mallien käyttöoikeuksia rajoittaa ja differentiaalista yksityisyyttä käyttää. Epämukava ydin on tämä: 'keskimäärin yksityinen' ei ole yksityisyystakuu, ja se pettää juuri ne potilaat, joilla on muutenkin vähiten suojaa.

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

Yksityisyystakuu on lupaus ihmiselle, ei keskiarvolle

Kun lääketieteellistä tekoälymallia kutsutaan ”yksityisyyttä säilyttäväksi”, väite perustuu yleensä yhteen lukuun: kaikkien koulutusdataan kuuluneiden potilaiden yli laskettuna keskimääräinen todennäköisyys päätellä yksittäisen henkilön jäsenyys on pieni. Se kuulostaa rauhoittavalta. Tämän tutkimuksen hiljainen ja epämukava huomio on, että se on väärä luku.

Yksityisyys ei ole keskiarvo. Se on jokaiselle ihmiselle annettu lupaus siitä, ettei koulutusdataan kuuluminen myöhemmin paljasta häntä. Keskiarvo voi pitää lupauksen lähes kaikille ja rikkoa sen täydellisesti muutaman kohdalla. Tutkijat mittasivat yksityisyysriskin potilas kerrallaan tarkkuudella, jota ei aiemmin ollut käytetty, ja löysivät juuri tämän: aggregaattina turvallisilta näyttävät mallit voivat vuotaa tiettyjen henkilöiden jäsenyyden lähes täydellisesti — ja suhteettoman usein juuri niiden, joilla on jo ennestään vähiten suojaa.

Mitä tutkijat tekivät

Moritz Knollen johtama ryhmä — mukana Daniel Rückert, molemmat Münchenin teknillisestä yliopistosta, Georg Kaissis Hasso Plattner Institutesta sekä tutkijoita Imperial College Londonista — otti tunnetun yksityisyysuhan, **membership inference -hyökkäyksen**, ja muutti kysymystä. Sen sijaan että olisi kysytty ”kuinka usein hyökkäys onnistuu keskimäärin koko aineistossa?”, kysyttiin ”kuinka altis *juuri tämä potilas* on?”

Potilaskohtainen analyysi tehtiin **seitsemässä vakiintuneessa lääketieteellisessä aineistossa**, jotka kattoivat hyvin erilaisia datatyyppejä: lääketieteellistä kuvantamista, elektrokardiogrammeja ja sähköisiä potilastietoja. Kussakin aineistossa analysoitiin 200 mallia. Jokaiselle potilaalle jokaisessa mallissa arvioitiin, kuinka varmasti hyökkääjä voisi päätellä, oliko henkilön tieto ollut koulutusdatassa. Tuloksia tarkasteltiin sen jälkeen ryhmittäin: sairausstatus, itse ilmoitettu rotu, vakuutus, sukupuoli ja kuvantamisprotokolla.

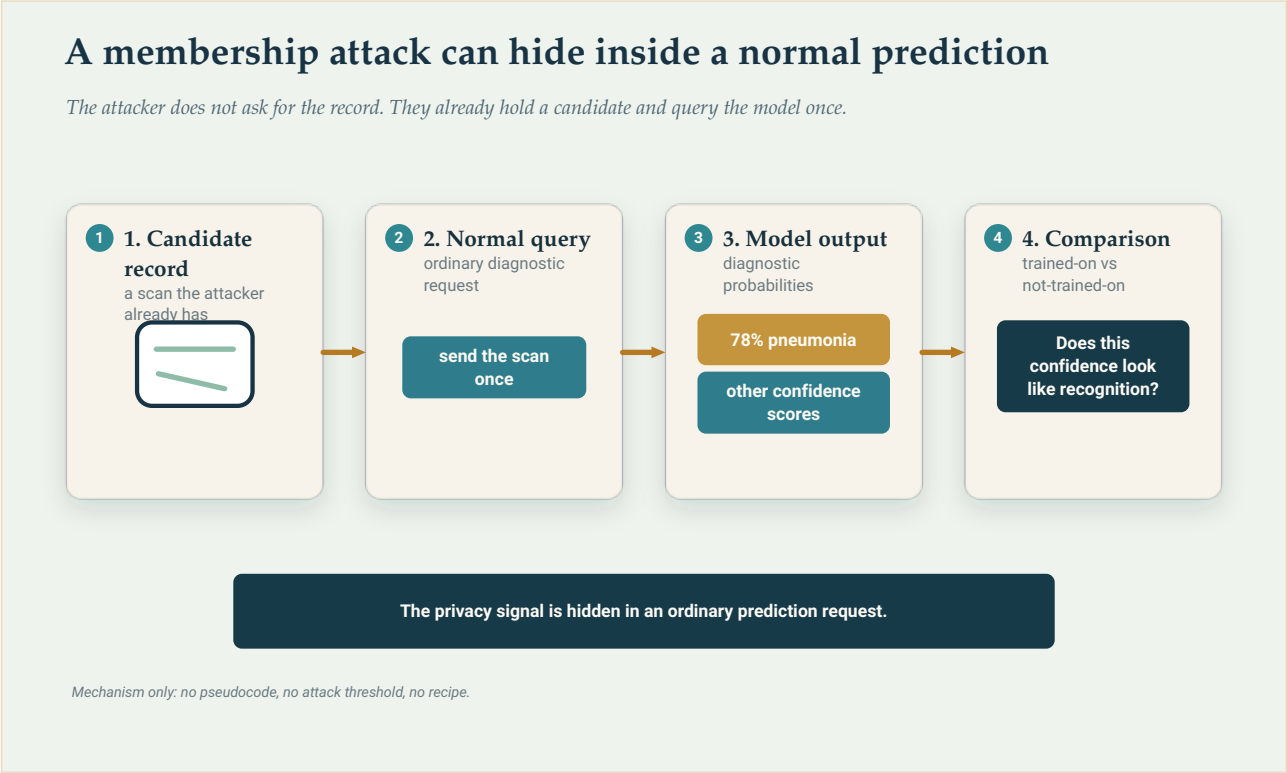
> **Mitä membership inference -hyökkäys oikeastaan on**
>
> Tässä vuoto on hienovaraisempi kuin ”malli tulostaa potilastietosi”. Se koskee *jäsenyyttä* — pelkkää tosiasiaa siitä, että tietosi oli koulutusaineistossa.
> Aloitetaan mallin tavallisesta toiminnasta. Syötät sille esimerkiksi rintakehän röntgenkuvan, ja se palauttaa todennäköisyyksiä: 78% keuhkokuumeelle sekä arviot kardiomegaliasta, ödeemasta ja konsolidaatiosta. Hyökkäys käyttää *vain* tätä tavallista diagnostista vastausta. Ei mitään eksoottista.
> Heikkous on tässä: malli on yleensä hieman **varmempi** juuri niistä esimerkeistä, joilla sitä koulutettiin, kuin sellaisista joita se ei ole koskaan nähnyt. Ajattele opiskelijaa, joka näki kokeen salaa etukäteen — harjoitelluissa kysymyksissä vastaus tulee hieman liian nopeasti ja itsevarmasti. Sen päättelemistä tämän vihjeen perusteella, kuuluiko tietty tieto mallin opiskeluesimerkkeihin, kutsutaan membership inferenceksi.

Hyökkäys etenee näin. Joku haluaa tietää, käytettiinkö tietyn henkilön kuvaa mallin koulutukseen. Hyökkääjä **ei** pyydä mallia palauttamaan tietoa — hänellä on jo ehdokaskuva, henkilön oma tai hyvin läheinen kopio. Se lähetetään mallille kerran tavallisena diagnostisena kyselynä ja vastauksen varmuus kirjataan. Sitten tarkistetaan, muistuttaako varmuus enemmän mallia, joka *oli* koulutettu tällä kuvalla, vai mallia joka *ei ollut*. Vertailu voidaan tehdä suhteellisen helposti kouluttamalla oma sijaismalli eli ”reference model” oppimaan, miltä tällainen paljastava ylivarmuus näyttää tavallisella tietokoneella ilman erityislaitteistoa. Jos oikea malli on epätavallisen varma juuri tästä kuvasta — ikään kuin tunnistaisi sen — se on vahvaa näyttöä koulutusjäsenyydestä.

Häiritsevää on, ettei kysely näytä hyökkäykseltä. Se on sama diagnostinen kysely, jonka kliinikko tekisi, ja yksityisyyssignaali kätkeytyy tavallisen ennusteen sisään. Siksi riski on konkreettinen, vaikka se on toistaiseksi osoitettu määritellyissä laboratorio-oletuksissa eikä havaittu luonnossa.

Miksi jäsenyydellä on väliä, jos itse tietue ei vuoda? Koska *jäsenyys on itsessään tieto*. Jos malli koulutettiin tietyn syövän immunoterapiaa saaneilla potilailla, sen vahvistaminen että sinun tietosi kuuluu aineistoon kertoo, että sinulla todennäköisesti oli kyseinen syöpä — juuri sellaista, mitä vakuutusyhtiön tai työnantajan ei pitäisi pystyä päättelemään. Sisältö pysyy suljettuna; kuulumisen tosiasia vuotaa.

Tutkijat kuvaavat oletukset avoimesti — pääsy mallin tavallisiin ennusteisiin, ehdokastieto ja hyökkääjän oma vertailumalli — eivät toimintaohjeena vaan jotta uhkaa voidaan arvioida eikä vain huitoa sivuun.



Hyökkäys ei tarvitse erityistä yksityisyysrajapintaa. Tavallinen diagnostinen kysely voi vuotaa jäsenyyssignaalin, jos malli on epätavallisen varma aiemmin näkemästään tiedosta.

Mitä he löysivät

Kolme havaintoa, joista jokainen on edellistä terävämpi.

Keskiarvot piilottavat altistuneet. Tavalliseen tapaan aggregaattina mitattuna monet mallit näyttävät rauhoittavan yksityisiltä: hyökkäys ei useimmilla potilailla menesty sattumaa paremmin. Potilaskohtainen näkymä kertoi toisen tarinan. Kuten Knolle sanoi tutkimuksen tiedotteessa, aiemmissa arvioissa ”on mitattu vain keskimääräinen riski kaikkien potilaiden yli. Me tutkimme riskiä ensimmäistä kertaa yksittäisen potilaan tasolla — ja kuva näyttää hyvin erilaiselta.” Joillakin yksilöillä hyökkäys onnistuu **lähes täydellisesti**: attack AUC on **0.95 tai enemmän** (0.5 vastaa kolikonheittoa, 1.0 täydellistä erottelua), vaikka koko aineiston keskiarvo näyttäisi sattuman tasolta.

Keskiarvo voi olla lähellä sattumaa samalla kun pieni tietueiden häntä on paljon alttiimpi tunnistamiselle. Tutkimuksen ydin on, että yksityisyys pitää tarkistaa potilastasolla, ei vain aggregaattina.

Altistuminen on epätasaista — ja kohdistuu aliedustettuihin. Lähes täydelliselle hyökkäykselle haavoittuvimmat potilaat kuuluivat järjestelmällisesti datassa **aliedustettuihin ryhmiin**: etnisiin vähemmistöihin, harvinaisten sairauksien fenotyyppeihin ja epätavallisiin kuvantamispiirteisiin. Sähköisten potilastietojen aineistossa mustat potilaat esiintyivät kaikkein haavoittuvimpien tietueiden joukossa **31% useammin** kuin odotettiin. Mammografia-aineistossa maligniteetin suhteen epäilyttäviksi merkittyjä kuvia oli tässä vaarahännässä peräti **+1,179%** enemmän kuin odotettiin.

Nämä kaksi lukua ovat esimerkkejä eivätkä koko kuva; tutkimus raportoi samanlaisen vinouman useissa ryhmissä. Ne ovat *suhteellisia yliedustuksia* pienessä ääririskin hännässä: kuinka paljon useammin nämä potilaat esiintyvät kaikkein altistuneimpien tietueiden joukossa, eivät sitä osuutta kaikista mustista tai epäilyttävän mammografian potilaista, joka olisi altistunut. Kun mallilla on vähemmän samankaltaisia esimerkkejä, tällaiset potilaat erottuvat — ja juuri erottumista membership-hyökkäys hyödyntää. Yksityisyysvirhe ei ole satunnainen, vaan kasaantuu jo valmiiksi marginaaliin jääville ihmisille.

Suuremmat mallit pahentavat ongelmaa. Lähes täydellisille hyökkäyksille altistuvien potilaiden määrä kasvoi jyrkästi mallikapasiteetin mukana. Yhdessä dermatologia-aineistossa osuus nousi käytännössä nollasta pienimmässä mallissa noin **yhteen kymmenestä** suurimmassa. Kun lääketieteelliset tekoälymallit kasvavat suuremmiksi ja kyvykkäämmiksi — juuri siihen suuntaan koko ala kulkee — tämä tietty riski pahenee, ei lievene.

Rückertin yhteenveto on suora: ”Tämä ei ole hyväksyttävä riski. Terveystiedot ovat erittäin arkaluonteisia.”

Miksi ”keskimäärin turvallinen” on väärä testi

Alhainen keskimääräinen riski on helppo tulkita puhtaaksi terveystodistukseksi. Tutkimuksen ydinoivallus on, ettei keskiarvo ole tässä vain epätarkka — se mittaa väärää asiaa.

Yksityisyystakuu on mielekäs vain, jos se pitää eniten vaarassa olevan henkilön kohdalla, ei mediaanihenkilön. Malli, jossa 999 potilasta 1,000 potilaasta ei ole tunnistettavissa mutta yksi voidaan tunnistaa lähes täydellisesti, ei ole ”99.9% yksityinen” missään siinä mielessä, jolla on tälle yhdelle ihmiselle merkitystä. Jos juuri tämä henkilö on ennustettavasti harvinaissairas tai etniseen vähemmistöön kuuluva, mittari ei ole vain puutteellinen vaan hiljaisesti syrjivä: se raportoi enemmistön turvallisuuden ja kutsuu sitä kaikkien turvallisuudeksi.

Tutkimus pakottaa siirtämään kysymyksen muodosta *kuinka yksityinen malli on keskimäärin?* muotoon *kuka on kaikkein altistunein potilas, ja kuka hän on?* Ne ovat eri kysymyksiä, ja vain jälkimmäinen on varsinainen yksityisyyskysymys.

Mitä tämä *ei* todista — eikä väitä

- Se **ei** sano, että lääketieteellinen tekoäly pitäisi hylätä. Tutkijoiden kehys on riskin pienentäminen, ei vetäytyminen: mittaa ja korjaa riski, älä lopeta rakentamista.
- Se **ei** tarkoita, että jokainen lääketieteellinen malli vuotaisi tai että jotakin käytössä olevaa mallia olisi todella hyökätty. Se osoittaa, että *riski on olemassa ja jakautuu epätasaisesti* ja että tavalliset aggregaattimittarit jättävät sen huomaamatta.
- Se **ei** tarkoita, että potilastietosi olisivat jo paljastuneet. Hyökkäys vaatii tietyt olosuhteet: pääsyn malliin, ehdokastiedon ja hyökkääjän oman infrastruktuurin. Kyse ei ole satunnaisesta kyvystä.

- Se **ei** osoita potilastiedon *sisällön* vuotavan. Vuotava asia on jäsenyys eli mukanaolon tosiasia, joka on vaarallinen eri syystä eikä siksi, että tietue tulostuisi ulos.
- Se **ei** tiivistä eroja yhteen otsikkolukuun. Vahva ja toistuva tulos on *kuvio*: aggregaattimittarit aliarvioivat yksilöllistä riskiä ja aliedustetut potilaat kantavat suurimman osan siitä useissa aineistoissa ja sadoissa malleissa.

Kuinka vahvaa näyttö on?

Kyseessä on empiirinen menetelmällinen tulos, ja se on robusti. Kuvio — aggregoidut yksityisyysmittarit aliarvioivat järjestelmällisesti potilaskohtaista riskiä, jäljelle jäävä riski keskittyy aliedustettuihin ryhmiin ja pahenee mallikapasiteetin kasvaessa — toistui **seitsemässä erilaisia datatyyppejä sisältävässä aineistossa** ja suuressa määrässä malleja per aineisto. Juuri tämä leveys tekee mittausväitteestä uskottavan yhden aineiston artefaktin sijasta.

Kaksi rehellistä varausta. Ensinnäkin tutkimus osoittaa *riskin* ajamalla tutkijoiden rakentamia hyökkäyksiä. Se luonnehtii, kuinka hyvin kyvykäs hyökkääjä *voisi* onnistua annetuilla oletuksilla, ei sitä kuinka usein tällaisia hyökkäyksiä todella tapahtuu. Toiseksi näyttävät luvut — joidenkin potilaiden lähes täydellinen tunnistus — kuvaavat tarkoituksella kaikkein huono-osaisimpia yksilöitä. Juuri se on tutkimuksen pointti, ja luvut pitää lukea muotoon ”häntä on paljon raskaampi kuin keskiarvo antaa ymmärtää”, ei ”useimmat potilaat ovat altistuneita”.

Oikea asenne ei ole hälytys eikä vähättely: huolellinen monen aineiston tutkimus osoittaa, että tavallinen tapa sertifioida lääketieteellisen tekoälyn yksityisyyttä on sokea omille pahimmille tapauksilleen — ja nämä tapaukset osuvat potilaisiin, joilla on vähiten pelivaraa.

Miksi tällä on merkitystä

Kaksi asiaa tekee tästä enemmän kuin teknisen sivuhuomautuksen.

Ensinnäkin se muuttaa sitä, mitä ”yksityisyyttä säilyttävän” pitäisi saada tarkoittaa. Jos malli julkaistaan keskimääräisen yksityisyyspisteen kanssa, piste voi olla aidosti pieni ja silti kätkeä joukon potilaita, jotka membership-hyökkäys tunnistaa lähes täydellisesti. Tutkimuksen käytännön vaatimus on konkreettinen: arvioi yksityisyysriski **potilaskohtaisesti** ennen julkaisua, kontrolloi kuka pääsee käyttämään malleja ja käytä **differentiaalisen yksityisyyden** kaltaisia tekniikoita — pieniä, tarkasti kalibroituja kohinalisäyksiä koulutuksessa, jotka heikentävät membership-hyökkäyksiä todellisella ja hallitulla kustannuksella mallin hyödyllisyydelle. Yksityisyys–hyöty-kompromissi on avoin, ei ilmainen. ”Tarkistimme keskiarvon” ei saisi enää tarkoittaa, että tarkistus on tehty.

Toiseksi tutkimus punoo yksityisyyden ja oikeudenmukaisuuden yhteen. Samat ryhmät, jotka ovat lääketieteellisessä datassa aliedustettuja — ja joita tekoäly siksi jo palvelee huonommin — osoittautuvat ryhmiksi, joiden yksityisyyttä tekoäly suojaa heikoimmin. Ala, joka yrittää korjata ensimmäistä epätasa-arvoa, ei voi käsitellä toista jonkun muun osastona. Kyse ovat samoista ihmisistä.

Rauhoittava tarina lääketieteellisen tekoälyn yksityisyydestä — *mittasimme sen, keskiarvo on pieni, kaikki hyvin* — on juuri se tarina, jonka tutkimus purkaa. Ei pelotellakseen ihmisiä pois teknologiasta, vaan siirtääkseen standardin sinne minne se kuuluu: lupaus voidaan väittää pidetyksi vasta, kun se on tarkistettu sen henkilön kohdalla, jota rikkominen todennäköisimmin satuttaa.

Lyhyesti

Membership inference -hyökkäykset yrittävät päätellä, oliko tietyn henkilön tieto tekoälymallin koulutusaineistossa. Koska jäsenyys voi itsessään paljastaa arkaluonteisen asian — esimerkiksi että henkilöllä oli tietty sairaus — se on todellinen yksityisyysvuoto, vaikka itse tietue ei koskaan tulisi näkyviin. Aiemmissa töissä mitattiin, kuinka usein hyökkäykset onnistuvat *keskimäärin* aineiston yli. Tämä tutkimus mittasi niitä *potilaskohtaisesti* seitsemässä lääketieteellisessä aineistossa (kuvantaminen, EKG, sähköiset tiedot) ja monissa malleissa. Aggregaattimittarit aliarvioivat riskin pahasti: jotkin yksilöt voidaan tunnistaa lähes täydellisesti (attack AUC $\geq$ 0.95), vaikka koko aineiston keskiarvo näyttää turvalliselta; haavoittuvimmat kuuluvat järjestelmällisesti aliedustettuihin ryhmiin, kuten etnisiin vähemmistöihin, harvinaissairaisiin ja poikkeavan kuvantamisen potilaisiin; ja ongelma kasvaa mallikoon mukana. Tutkijat eivät ehdota lääketieteellisen tekoälyn hylkäämistä, vaan yksilötason yksityisyysmittausta, mallien käyttöoikeuksien hallintaa ja differentiaalista yksityisyyttä. Johtopäätös: ”keskimäärin yksityinen” ei ole yksityisyystakuu, ja sen pettämät ihmiset ovat yleensä niitä, joilla on jo vähiten suojaa.

Ei-hölynpölyä-tarkistus

Mitä tutkimus osoittaa: Seitsemässä lääketieteellisessä aineistossa ja monissa malleissa potilaskohtainen membership inference -riski on joillekin yksilöille paljon suurempi kuin aggregaattimittarit vihjaavat — jopa lähes täydellinen hyökkäysmenestys (AUC $\geq$ 0.95) — ja jäljelle jäävä riski kohdistuu suhteettomasti aliedustettuihin ryhmiin sekä kasvaa mallikapasiteetin mukana.

Mikä on uskottavaa mutta todistamatta: Että reaalimaailman hyökkääjät käyttävät tätä jo käytössä oleviin kliinisiin malleihin. Tutkimus osoittaa *kyvyn ja sen jakautumisen* annetuilla oletuksilla, ei hyökkäysten yleisyyttä luonnossa.

Mitä tutkimus ei osoita: Että lääketieteellinen tekoäly pitäisi hylätä; että jokainen malli vuotaa tai että jokin tietty käytössä oleva malli on murrettu; että potilastietueen *sisältö* jäsenyyden sijasta vuotaa; yhtä yksittäistä eriarvoisuuslukua — robusti tulos on kuvio, ei yksi numero.

Tärkeimmät rajoitukset: Tutkimus mittaa pahimman tapauksen riskiä tutkijoiden omien hyökkäysten kautta eikä havaittuja tapauksia; näyttävät luvut kuvaavat tarkoituksella kaikkein altistuneimpia yksilöitä; tarkat ryhmäerot näkyvät johdonmukaisena kuviona eri aineistoissa eikä niitä tiivistetä yhteen lukuun.

Kuinka paljon yleislukijan kannattaa luottaa tulokseen? Korkealla varmuudella aggregoidut yksity-

isyysmittarit aliarvioivat yksilöllistä riskiä, jäljelle jäävä riski keskittyy aliedustettuihin potilaisiin ja ongelma pahenee mallikoon kasvaessa — tämä näkyy monissa aineistoissa ja malleissa. Kohtalaisella varmuudella reaalimaailman hyökkäysten yleisyyteen, jota tutkimus ei mittaa. Turvallinen tulkinta ei ole ”lääketieteellinen tekoäly vuotaa datasi” eikä ”yksityisyys on ratkaistu”, vaan ”tavallinen yksityisyystarkistus on sokea omille pahimmille tapauksilleen, ja ne osuvat haavoittuvimpiin potilaisiin — joten tarkistuksen pitää muuttua”.

Lähteet

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