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

'n Mediese KI kan gemiddeld privaat wees en steeds spesifieke pasiënte blootstel — veral die onderverteenwoordigdes

'n Mediese-KI-model wat "privaatheidsbewarend" genoem word, berus gewoonlik op een gemiddelde getal. Hierdie studie voer aan dat dit die verkeerde toets is. Die outeurs het lidmaatskapafleidingsaanvalle bestudeer — aanvalle wat verklap of 'n spesifieke persoon se rekord in 'n model se opleidingsdata was en dus kan verklap dat hulle 'n bepaalde siekte gehad het — en risiko per pasiënt eerder as in die geheel gemeet, oor sewe mediese datastelle (beeldvorming, EKG, elektroniese rekords) en baie modelle elk. Die patroon: modelle wat gemiddeld veilig lyk, kan 'n aanvaller steeds toelaat om spesifieke individue byna perfek te identifiseer (aanval-AUC $\geq$ 0,95); die blootgesteldes kom sistematies uit onderverteenwoordigde groepe (etniese minderhede, seldsame siekte, ongewone beeldvorming); en dit word erger namate modelle groei. Die outeurs sê nie mediese KI moet laat vaar word nie — hulle sê meet privaatheid per pasiënt, beheer modeltoegang en gebruik differensiële privaatheid. Die ongemaklike kern: "gemiddeld privaat" is nie 'n privaatheidswaarborg nie, en dit faal die pasiënte wat reeds die minste beskerm is.

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

’n Privaatheidswaarborg is ’n belofte aan ’n persoon, nie aan ’n gemiddelde nie

Wanneer ’n mediese-KI-model “privaatheidsbewarend” genoem word, berus dié aanspraak gewoonlik op een getal: oor al die pasiënte wie se data die model opgelei het, is die gemiddelde kans laag dat iemand kan aflei of ’n spesifieke persoon deel van die opleidingsdata was. Dit klink gerusstellend. Hierdie artikel se stil, ongemaklike punt is dat dit die verkeerde getal is.

Privaatheid is nie ’n gemiddelde nie. Dit is ’n belofte aan elke individu — dat die feit dat hulle in die opleidingsdata was nie later teen hulle gebruik kan word nie. En ’n gemiddelde kan daardie belofte vir byna almal hou terwyl dit vir ’n paar mense volledig breek. Die navorsers het privaatheidsrisiko een pasiënt op ’n slag gemeet, met ’n resolusie wat nog nie voorheen gebruik is nie, en presies dit gevind: modelle wat in die geheel veilig lyk, kan die lidmaatskap van spesifieke individue byna perfek verklap — en die individue wat blootgestel word, is buite verhouding dié wat reeds die minste beskerm is.

Wat die outeurs gedoen het

Die span — gelei deur Moritz Knolle, saam met Daniel Rückert (albei by die Technical University of Munich) en Georg Kaissis (Hasso Plattner Institute), met kollegas by Imperial College London — het ’n bekende privaatheidsbedreiging, ’n **lidmaatskap-afleidingsaanval**, geneem en die vraag verander. In plaas daarvan om te vra “hoe dikwels slaag hierdie aanval gemiddeld oor die datastel?”, het hulle gevra “hoe blootgestel is *hierdie spesifieke pasiënt*?”

Hulle het dié per-pasiënt-ontleding oor **sewe gevestigde mediese datastelle** uitgevoer, oor baie verskillende datatipes — mediese beeldvorming, elektrokardiogramme en elektroniese gesondheidsrekords — en, oor die datastelle heen, 200 modelle elk. Vir elke pasiënt in elke model het hulle geraam hoe seker ’n aanvaller kon bepaal of daardie persoon se rekord deel van die opleidingsdata was, en die resultate daarna volgens groep opgebreek: siektestatus, selfgerapporteerde ras, versekering, geslag en beeldprotokol.

> **Wat ’n lidmaatskap-afleidingsaanval eintlik is**
>
> Die lekkasie hier is subtieler as “die model spoeg jou rekord uit”. Dit gaan oor *lidmaatskap* — bloot die feit dat jou data in die opleidingsstel was.
> Begin by wat een van hierdie modelle normaalweg doen. Jy gee dit ’n skandering — byvoorbeeld ’n borskas-X-straal — en dit gee waarskynlikhede terug: ’n 78%-kans op longontsteking, saam met lesings vir kardiomegalie, edeem en konsolidasie. Daardie alledaagse diagnostiese antwoord is die *enigste* ding wat die aanval gebruik. Niks eksoties nie.
> Die swakheid waarop dit steun, is hierdie: ’n model is gewoonlik ’n bietjie **meer seker** oor die presiese voorbeelde waarop dit opgelei is as oor voorbeelde wat dit nog nooit gesien het nie. Dink aan ’n student wat die eksamen vooraf in die geheim gesien het — op die vrae wat reeds geoefen is, kom die antwoorde ’n bietjie te vinnig en ’n bietjie te selfversekerd. Om uit dié teken

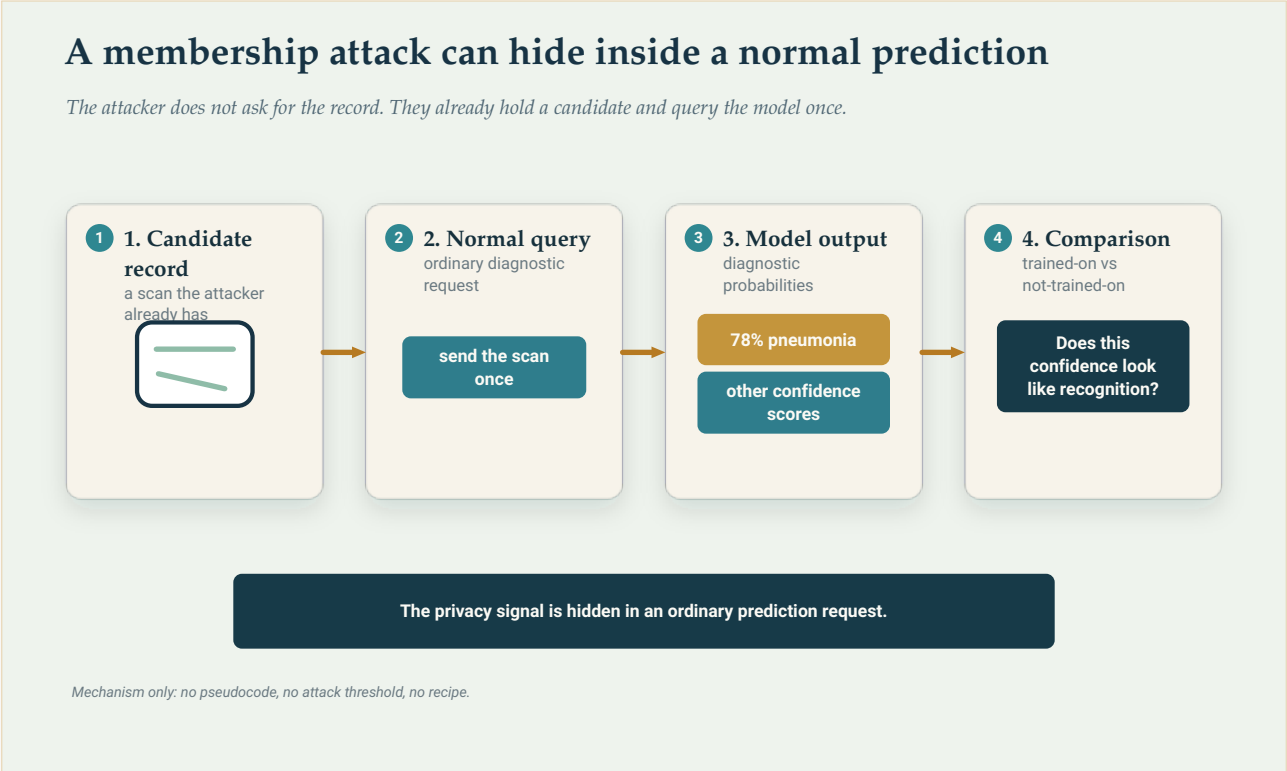
af te lei of ’n spesifieke rekord een van die voorbeelde was wat die model bestudeer het, is wat “lidmaatskap-afleiding” beteken.

Die aanval werk dus stap vir stap so. Iemand wil weet of ’n spesifieke persoon se skandering gebruik is om die model op te lei. Hulle vra die model **nie** vir die rekord nie — hulle het reeds ’n kandidaat-skandering (die persoon s’n, of ’n baie soortgelyke kopie). Hulle stuur dit een keer in as ’n gewone diagnostiese versoek en let op hoe seker die antwoord is. Daarna vergelyk hulle daardie sekerheid met hoe ’n model lyk wat *wel* op die skandering opgelei is en een wat *nie* was nie — ’n vergelyking wat goedkoop gedoen kan word deur ’n eie plaasvervangermodel (“verwysingsmodel”) op te lei om dié kenmerkende oorvertroue te leer, op ’n gewone rekenaar sonder spesiale hardeware. As die werklike model buitengewoon seker is oor hierdie skandering — asof dit dit herken — is dit sterk bewys dat die skandering in die opleidingsdata was.

Die ontstellende deel is dat niks aan die versoek soos ’n aanval lyk nie: dit is dieselfde diagnostiese navraag wat ’n klinikus sou maak, en die privaatheidsein is binne ’n gewone voorspelling versteek. Daarom is die risiko konkreet — al is dit tot dusver onder verklaarde laboratoriumaannames gedemonstreer eerder as in die natuur vasgevang.

Waarom maak lidmaatskap saak as die rekord self nooit uitlek nie? Omdat *lidmaatskap ’n feit is*. As ’n model opgelei is op pasiënte wat ’n bepaalde kanker-immunoterapie ontvang het, kan bevestiging dat jou rekord daarin is verklap dat jy waarskynlik daardie kanker gehad het — presies die soort ding wat ’n versekeraar of werkgewer nooit behoort te kan aflei nie. Die inhoud bly verseël; dit is die feit dat jy tot die groep behoort wat uitlek.

Die outeurs sit dié aannames openlik uiteen — toegang tot die model se gewone voorspellings, ’n kandidaatrekord en die aanvaller se eie verwysingsmodel — nie as ’n handleiding nie, maar sodat oor die bedreiging geredeneer kan word eerder as om dit weg te wuif.



Die aanval het nie 'n spesiale privaatheids-API nodig nie. 'n Gewone diagnostiese navraag kan 'n lidmaatskapsein verklap as die model buitengewoon seker is oor 'n rekord wat dit al voorheen gesien het.

Wat hulle gevind het

Drie bevindings, elk skerper as die vorige een.

Gemiddeldes versteek die blootgesteldes. In die geheel gemeet — die gewone manier — lyk baie van hierdie modelle gerusstellend privaat: vir die meeste pasiënte vaar die aanval nie beter as kans nie. Die per-pasiënt-prentjie het 'n ander verhaal vertel. Soos Knolle dit in die studie se aankondiging gestel het, het vorige assesserings "nog altyd net die gemiddelde risiko oor alle pasiënte gemeet. Ons het vir die eerste keer die risiko op die vlak van individuele pasiënte ondersoek — en dit skets 'n heel ander prentjie." Vir sommige individue slaag die aanval **byna perfek** — die outeurs meet dit as 'n aanval-AUC van **0,95 of hoër** (0,5 is 'n muntgooi, 1,0 is foutloos) — selfs wanneer die gemiddelde oor die hele datastel nie beter as kans lyk nie.

Die gemiddelde kan naby kansvlak lê terwyl 'n klein stert van rekords baie meer blootgestel bly. Die artikel se punt is dat privaatheid op pasiëntvlak nagegaan moet word, nie net in die geheel nie.

Original Aurora diagram — The Clean Paper CC BY 4.0

Die blootstelling is ongelyk — en tref die onderverteenwoordigdes. Die pasiënte wat die kwesbaarste vir 'n byna perfekte aanval was, was sistematies dié in groepe wat **onderverteenwoordig in die data** was: etniese minderhede, seldsame-siekte-fenotipes en ongewone beeldkenmerke. In die elektroniese-rekorddatastel het swart pasiënte **31% meer dikwels** as verwag onder die mees kwesbare rekords verskyn; in die mammografiestel was skanderings wat as verdag vir maligniteit gemerk is met 'n opvallende **+1 179%** oorverteenwoordig in daardie gevaarsone. (Dié twee syfers is voorbeelde, nie die hele prentjie nie — die artikel rapporteer dieselfde skeeftrekking oor verskeie groepe — en dit is *relatiewe* oorverteenwoordiging binne die klein uiterste-risiko-stert: hoeveel meer gereeld dié pasiënte onder die mees blootgestelde rekords voorkom, nie die aandeel swart pasiënte of verdagte mammogramme wat blootgestel is nie.) 'n Model het minder soortgelyke voorbeelde waarin dié pasiënte kan "wegraak", dus staan hul rekords uit — en om uit te staan is presies wat 'n lidmaatskapaanval opspoor. Die privaatheidsmislukking is nie lukraak nie; dit konsentreer op mense wat reeds op die marges is.

Groter modelle maak dit erger. Die aantal pasiënte wat aan byna perfekte aanvalle blootgestel is, het skerp met modelkapasiteit gestyg — in een dermatologiedatastel het die aandeel van feitlik nul in die kleinste model tot ongeveer **een uit tien** in die grootste geklim. Namate mediese KI-modelle groter en kragtiger word — die rigting waarin die hele veld beweeg — word hierdie spesifieke risiko volgens die outeurs ernstiger, nie kleiner nie.

Rückert se opsomming is reguit: "Dit is nie 'n aanvaarbare risiko nie. Gesondheidsdata is hoogs sensitief."

Waarom “gemiddeld veilig” die verkeerde toets is

Die versoeking is om ’n lae gemiddelde risiko as ’n ondubbelsinnige gesondheidsverklaring te lees. Die kernles van die artikel is dat ’n gemiddelde hier nie net onpresies is nie — dit meet die verkeerde ding.

’n Privaatheidswaarborg beteken net iets as dit geld vir die persoon met die grootste risiko, nie net vir die middelste persoon nie. ’n Model waarin 999 uit 1 000 pasiënte nie geïdentifiseer kan word nie maar een byna perfek geïdentifiseer kan word, is nie “99,9% privaat” in enige sin wat vir daardie een persoon saak maak nie — en as daardie een persoon voorspelbaar die pasiënt met ’n seldsame siekte of die pasiënt uit ’n etniese minderheid is, is die maatstaf nie net onvolledig nie, maar stilweg diskriminerend. Dit rapporteer veiligheid vir die meerderheid en noem dit veiligheid vir almal.

Dit is die verskuiwing wat die artikel afdwing: van *hoe privaat is hierdie model gemiddeld?* na *wie is die mees blootgestelde pasiënt, en wie is daardie persoon?* Dit is verskillende vrae, en net die tweede een is werklik ’n privaatheidsvraag.

Wat dit *nie* bewys — of beweer — nie

- Dit sê **nie dat mediese KI laat vaar moet word nie**. Die outeurs se raamwerk is versagting, nie terugtog nie: meet en herstel die risiko; moenie ophou bou nie.
- Dit beteken **nie dat elke mediese model lek nie**, of dat enige spesifieke ontplooide model reeds aangeval is nie. Dit wys dat die *risiko bestaan en ongelyk versprei is*, en dat standaard-aggregaatmaatstawwe dit mis.
- Dit beteken **nie dat jou rekords reeds blootgestel is nie**. Die aanval vereis spesifieke voorwaardes — toegang tot die model, ’n kandidaatrekord en die aanvaller se eie infrastruktuur — nie bloot toevallige toegang nie.
- Dit wys **nie dat die inhoud van pasiëntrekords uitlek nie**. Wat uitlek is lidmaatskap — die feit van insluiting — wat om ’n ander rede gevaarlik is, nie omdat die rekord uitgegooi word nie.
- Dit **verminder nie die ongelykhede tot een opskrifgetal nie**. Die sterk, herhaalbare resultaat is die *patroon* — aggregaatmaatstawwe onderskat individuele risiko, en onderverteenwoordigde pasiënte dra die meeste daarvan — oor datastelle en honderde modelle.

Hoe sterk is die bewys?

Dit is ’n empiriese, metodologiese resultaat, en ’n robuuste een. Die patroon — aggregaat-privaatheidsmaatstawwe onderskat sistematies per-pasiënt-risiko, die oorblywende risiko konsentreer op onderverteenwoordigde groepe en dit vererger met modelkapasiteit — het oor **sewe datastelle van verskillende datatipes** en ’n groot aantal modelle per datastel standgehou. Daardie breedte is presies wat ’n meetaanspraak geloofwaardig maak eerder as ’n artefak van een datastel.

Twee eerlike voorbehoude. Eerstens is dit ’n demonstrasie van *risiko*, gemeet deur die aanvalle uit te voer wat die outeurs self gebou het; dit beskryf hoe goed ’n bekwame aanvaller *sou kon* vaar onder verklaarde aannames, nie hoe dikwels werklike aanvalle plaasvind nie. Tweedens beskryf die opvallende getalle — byna perfekte aanvalssukses vir sommige pasiënte — doelbewus die individue wat die slegste daaraan toe is; dit is die hele punt, en moet gelees word as “die stert is baie swaarder as wat die gemiddelde impliseer”, nie as “die meeste pasiënte is blootgestel” nie.

Die gepaste houding is nóg alarm nóg afwysing: ’n versigtige, multi-datastel-studie wat wys dat die standaardmanier waarop ons mediese-KI-privaatheid sertifiseer blind is vir sy eie ergste gevalle — en dat dié gevalle val op pasiënte met die kleinste veiligheidsmarge.

Waarom dit saak maak

Twee dinge maak dit meer as ’n tegniese voetnota.

Eerstens verander dit wat “privaatheidsbewarend” toegelaat behoort te beteken. As ’n model met ’n gemiddelde privaatheidstelling vrygestel word, kan dié telling werklik laag wees en steeds ’n subgroep pasiënte versteek wat byna perfek deur ’n lidmaatskapaanval geïdentifiseer kan word. Die artikel se praktiese eis is konkreet: beoordeel privaatheidsrisiko **per pasiënt** vóór vrystelling, beheer wie toegang tot ontplooide modelle kry en gebruik tegnieke soos **differensiële privaatheid** — klein, noukeurig gekalibreerde geraas wat tydens opleiding bygevoeg word en lidmaatskapaanvalle afstomp, teen ’n werklike en bestuurde koste aan die model se bruikbaarheid (die privaatheid–bruikbaarheid-afruil is uitdruklik, nie gratis nie). “Ons het die gemiddelde nagegaan” behoort nie meer as ’n volledige privaatheidskontrole te tel nie.

Tweedens vleg dit privaatheid en billikheid saam. Dieselfde groepe wat in mediese data onderverteenwoordig is — en daarom reeds swakker deur mediese KI bedien word — blyk ook dié te wees wie se privaatheid die KI die swakste beskerm. ’n Veld wat hard aan die eerste ongelykheid werk, kan die tweede nie as iemand anders se departement behandel nie. Dit is dieselfde mense.

Die gerusstellende verhaal oor mediese-KI-privaatheid — *ons het dit gemeet, die gemiddelde is laag, ons is veilig* — is die verhaal wat hierdie artikel uitmekaar haal. Nie om mense van die tegnologie af te skrik nie, maar om die standaard te skuif na waar dit hoort: ’n belofte wat jy net kan beweer jy hou as jy dit nagegaan het vir die persoon wat die waarskynlikste skade kan ly.

Kortliks

Lidmaatskap-afleidingsaanvalle probeer bepaal of ’n spesifieke persoon se rekord in ’n KI-model se opleidingsdata was — en omdat lidmaatskap self iets kan verklap (byvoorbeeld dat jy ’n bepaalde siekte gehad het), is dit ’n werklike privaatheidslek selfs wanneer die onderliggende rekord nooit verskyn nie. Vroeëre werk het gemeet hoe dikwels sulke aanvalle *gemiddeld* oor ’n datastel slaag. Hierdie studie het dit *per pasiënt* gemeet, oor sewe mediese datastelle (beeldvorming, EKG, elektroniese reko-

rds) en baie modelle elk, en gevind dat aggregaatmaatstawwe die risiko ernstig onderskat: sommige individue kan byna perfek geïdentifiseer word (aanval-AUC $\geq$ 0,95) selfs wanneer die gemiddelde oor die datastel veilig lyk; die kwesbaarstes is sistematies dié uit onderverteenwoordigde groepe (etniese minderhede, seldsame siektes, ongewone beeldvorming); en die probleem groei met modelgrootte. Die outeurs pleit nie vir die beëindiging van mediese KI nie — hulle pleit vir privaatheidsmeting op individuele vlak, beheer oor modeltoegang en die gebruik van differensiële privaatheid. Die kern: "gemiddeld privaat" is nie 'n privaatheidswaarborg nie, en dit faal juis die pasiënte wat reeds die minste beskerm is.

Nugtere beoordeling

Wat die artikel wys: Oor sewe mediese datastelle en baie modelle elk is per-pasiënt-risiko vir lidmaatskap-afleiding vir sommige individue veel hoër as wat aggregaatmaatstawwe suggereer — tot byna perfekte aanvalssukses (AUC $\geq$ 0,95) — en dié oorblywende risiko tref onderverteenwoordigde groepe buite verhouding en groei met modelkapasiteit.

Wat aannemelik is maar nie bewys nie: Dat werklike aanvallers dit reeds teen ontplooide kliniese modelle uitbuit; die studie demonstreer die *vermoë en die verspreiding daarvan* onder verklaarde aannames, nie die frekwensie van aanvalle in die natuur nie.

Wat dit nie wys nie: Dat mediese KI laat vaar moet word; dat elke model lek of enige spesifieke ontplooide model gekompromitteer is; dat die *inhoud* van pasiëntrekords (eerder as lidmaatskap) blootgestel word; een enkele gekwantifiseerde ongelykheidsyfer — die robuuste resultaat is die patroon, nie een getal nie.

Belangrikste beperkings: Dit meet ergste-geval-risiko via die outeurs se eie aanvalle, nie waargenome voorvalle nie; die dramatiese syfers beskryf doelbewus die mees blootgestelde individue; en die presiese groepsverskille word as 'n konsekwente patroon oor datastelle getoon eerder as tot een getal gereduseer.

Hoeveel vertroue behoort 'n algemene leser hierin te hê? Hoë vertroue dat aggregaat-privaatheidsmaatstawwe individuele risiko onderskat, dat die oorblywende risiko op onderverteenwoordigde pasiënte konsentreer en dat dit met modelgrootte vererger — dit word oor baie datastelle en modelle gedemonstreer. Matige vertroue oor die werklike frekwensie van sulke aanvalle, wat hierdie studie nie meet nie. Die veilige lees is nie "mediese KI lek jou data" of "privaatheid is opgelos" nie, maar "die standaard-privaatheidskontrole is blind vir sy eie ergste gevalle, en daardie gevalle val op die kwesbaarste pasiënte — dus moet die kontrole verander."

Bronne

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