
The Clean Paper

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

# Een medische AI kan gemiddeld privé zijn en toch specifieke patiënten blootstellen — vooral de ondervertegenwoordigde

*Een medisch AI-model dat "privacybeschermend" heet, rust meestal op één gemiddelde score. Deze studie betoogt dat dit de verkeerde test is. Bij membership-inference-aanvallen — die onthullen of het dossier van een specifieke persoon in de trainingsdata zat en daarmee bijvoorbeeld een ziekte kunnen verraden — maten de auteurs het risico per patiënt in plaats van geaggregeerd, over zeven medische datasets en veel modellen. Het patroon: modellen die gemiddeld veilig lijken kunnen specifieke personen bijna perfect identificeerbaar maken (aanval-AUC $\geq$ 0,95); de blootgestelden komen systematisch uit ondervertegenwoordigde groepen; en het wordt erger naarmate modellen groter worden. De auteurs zeggen niet dat medische AI moet worden opgegeven, maar dat privacy per patiënt moet worden gemeten, modeltoegang beperkt en differential privacy gebruikt.*

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

# Een privacygarantie is een belofte aan een persoon, niet aan een gemiddelde

Wanneer een medisch AI-model “privacybeschermend” wordt genoemd, rust die claim meestal op één getal: gemiddeld over alle patiënten van wie de gegevens het model hebben getraind, is de kans klein dat iemand kan afleiden of een specifieke persoon in de trainingsset zat. Dat klinkt geruststellend. Het stille, ongemakkelijke punt van dit artikel is dat dit het verkeerde getal is.

Privacy is geen gemiddelde. Het is een belofte aan ieder individu — dat deelname aan de trainingsset later niet tegen hem of haar kan worden gebruikt. Een gemiddelde kan die belofte voor bijna iedereen houden en haar voor een paar mensen volledig breken. De onderzoekers maten privacyrisico patiënt voor patiënt, op een resolutie die hiervoor nog niet was gebruikt, en vonden precies dat: modellen die op groepsniveau veilig lijken, kunnen voor specifieke personen bijna perfect verraden of hun gegevens in de training zaten — en die personen behoren onevenredig vaak tot de groepen die al het minst beschermd zijn.

## Wat de auteurs deden

Het team — onder leiding van Moritz Knolle, met Daniel Rückert (beiden Technical University of Munich) en Georg Kaissis (Hasso Plattner Institute), samen met collega’s van Imperial College London — nam een bekende privacydreiging, een **membership-inference-aanval**, en veranderde de vraag. Niet “hoe vaak slaagt deze aanval gemiddeld over de hele dataset?”, maar “hoe blootgesteld is *deze specifieke patiënt*?”

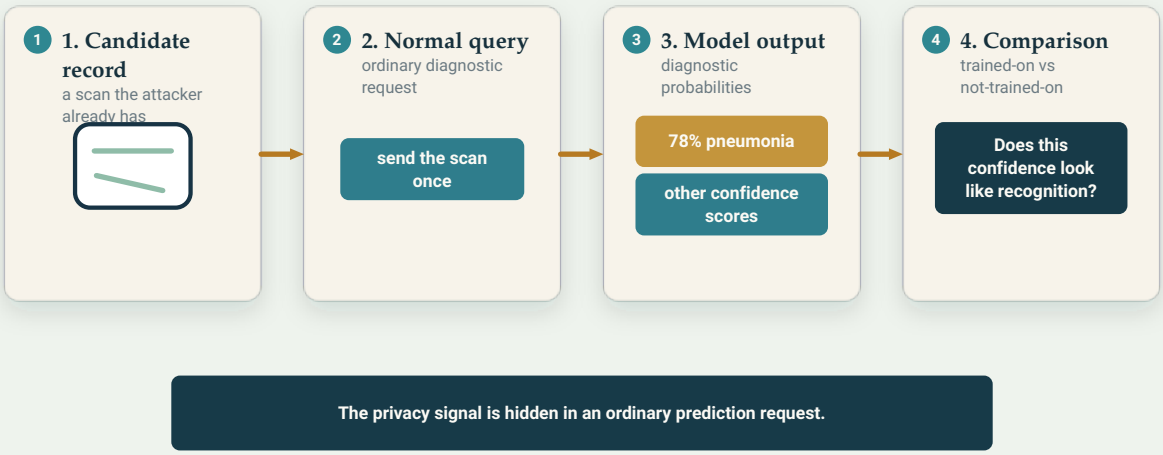
Ze voerden die analyse per patiënt uit op **zeven gevestigde medische datasets** met zeer verschillende gegevenstypen — medische beeldvorming, elektrocardiogrammen en elektronische patiëntendossiers — en per dataset op 200 modellen. Voor elke patiënt in elk model schatten ze hoe zeker een aanvaller kon bepalen of het dossier van die persoon deel van de trainingsdata was geweest. Daarna splitsten ze de resultaten uit naar groepen: ziektestatus, zelfgerapporteerde etniciteit/ras, verzekering, geslacht en beeldvormingsprotocol.

> **Wat een membership-inference-aanval eigenlijk is**
>
> Het lek is subtieler dan “het model spuugt je dossier uit”. Het gaat om *membership* — alleen al het feit dat jouw gegevens in de trainingsset zaten.
> Begin bij wat zo’n model normaal doet. Je geeft het bijvoorbeeld een thoraxfoto en krijgt waarschijnlijkheden terug: 78% kans op pneumonie, plus scores voor cardiomegalie, oedeem, consolidatie. Alleen dat gewone diagnostische antwoord gebruikt de aanval. Niets exotisch.
> De kwetsbaarheid waarop de aanval leunt is dat een model meestal een beetje **zekerder** is over precies de voorbeelden waarop het is getraind dan over voorbeelden die het nooit heeft gezien. Denk aan een leerling die stiekem het examen al kende: bij vragen die hij geoefend heeft komt het antwoord net iets sneller en zekerder. Uit dat signaal afleiden of een bepaald dossier

een van de voorbeelden was die het model heeft bestudeerd, is wat “membership inference” betekent.
Stap voor stap gaat het zo. Iemand wil weten of de scan van een bepaalde persoon voor training is gebruikt. De aanvaller vraagt het model **niet** om het dossier — hij beschikt al over een kandidaat-scan, bijvoorbeeld die van de persoon zelf of een zeer nabije kopie. Hij voert die één keer in als gewone diagnostische aanvraag en noteert hoe zeker het antwoord is. Daarna vergelijkt hij die zekerheid met wat je verwacht van een model dat de scan *wel* of *niet* in training zag. Dat patroon kan hij relatief goedkoop leren met een eigen referentiemodel op een gewone computer. Als het echte model ongewoon zeker is over deze scan — alsof het haar herkent — is dat bewijs dat de scan waarschijnlijk in de trainingsdata zat.
Het verontrustende is dat de aanvraag zelf niet op een aanval lijkt: het is dezelfde diagnostische query die een clinicus zou doen, en het privacysignaal zit verstopt in een gewone voorspelling. Precies daarom is het risico concreet, ook al is het tot nu toe onder expliciete laboratoriumaannames aangetoond en niet als incident in het wild waargenomen.
Waarom doet membership ertoe als het dossier zelf nooit uitlekt? Omdat *membership zelf een feit is*. Als een model is getraind op patiënten die een bepaalde kankerimmunotherapie kregen, kan bevestigen dat jouw dossier erin zat verraden dat je waarschijnlijk die kanker had — precies het soort informatie dat een verzekeraar of werkgever nooit zou mogen afleiden. De inhoud blijft afgesloten; het feit dat je tot de groep behoort ontsnapt.
De auteurs zetten deze aannames expliciet uiteen — toegang tot gewone modelvoorspellingen, een kandidaatdossier en een eigen referentiemodel — niet als handleiding, maar zodat over de dreiging concreet kan worden geredeneerd.

## 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.*

De aanval heeft geen speciale privacy-API nodig. Een gewone diagnostische query kan een membership-signaal lekken wanneer het model ongewoon zeker is over een dossier dat het eerder heeft gezien.



# Wat ze vonden

Drie bevindingen, elk scherper dan de vorige.

**Gemiddelden verbergen wie blootstaat.** Op geaggregeerd niveau — de gebruikelijke meting — zien veel modellen er geruststellend privé uit: voor de meeste patiënten doet de aanval het niet beter dan toeval. Per patiënt ontstond een ander beeld. Zoals Knolle het in de aankondiging van de studie formuleerde, maten eerdere beoordelingen “alleen het gemiddelde risico over alle patiënten”; hier werd voor het eerst op het niveau van individuele patiënten gekeken, met een heel ander beeld als gevolg. Voor sommige personen slaagt de aanval **bijna perfect** — door de auteurs gemeten als een aanval-AUC van **0,95 of hoger** (0,5 is muntgooien, 1,0 foutloos) — terwijl het gemiddelde van de hele dataset niet beter dan toeval lijkt.


Het gemiddelde kan rond toeval liggen terwijl een kleine staart van dossiers veel sterker blootstaat. Het punt van het artikel is dat privacy op patiëntniveau moet worden gecontroleerd, niet alleen geaggregeerd.



**De blootstelling is ongelijk — en treft ondervertegenwoordigde groepen.** Patiënten die het

kwetsbaarst waren voor bijna perfecte aanvallen kwamen systematisch uit groepen die **ondervertegenwoordigd waren in de data**: etnische minderheden, zeldzame ziektefenotypen, ongebruikelijke beeldkenmerken. In de dataset met elektronische dossiers kwamen zwarte patiënten **31% vaker** dan verwacht voor onder de kwetsbaarste dossiers; in de mammografiedataset waren scans die als verdacht voor maligniteit waren gemarkeerd in die gevarenzone maar liefst **+1.179%** oververtegenwoordigd. Deze twee cijfers zijn voorbeelden, niet het hele beeld — dezelfde scheefheid verschijnt in meerdere groepen — en het zijn *relatieve* oververtegenwoordigingen binnen de kleine staart met extreem hoog risico: hoeveel vaker zulke patiënten tussen de meest blootgestelde dossiers terechtkomen, niet welk percentage zwarte patiënten of verdachte mammografieën zelf is blootgesteld. Een model heeft minder vergelijkbare voorbeelden om zulke patiënten in te laten opgaan, waardoor hun dossiers eruit springen — precies wat een membership-aanval detecteert. De privacyfout is dus niet willekeurig; zij concentreert zich bij mensen die al aan de rand staan.

**Grotere modellen maken het erger.** Het aantal patiënten dat aan bijna perfecte aanvallen blootstaat nam sterk toe met de capaciteit van het model. In één dermatologiedataset steeg het aandeel van vrijwel nul bij het kleinste model naar ongeveer **één op tien** bij het grootste. Terwijl medische AI-modellen groter en krachtiger worden — precies de richting waarin het veld beweegt — wordt dit specifieke risico volgens de auteurs dus ernstiger, niet kleiner.

Rückerts samenvatting is kort: dit is geen aanvaardbaar risico; medische gegevens zijn uiterst gevoelig.

## Waarom “gemiddeld veilig” de verkeerde test is

De verleiding is om een laag gemiddeld risico als een vrijbrief te lezen. De kernles is dat middelen hier niet alleen onnauwkeurig is — het meet het verkeerde.

Een privacygarantie heeft alleen betekenis als zij geldt voor de persoon met het hoogste risico, niet voor de persoon in het midden. Een model waarin 999 van de 1.000 patiënten niet herkenbaar zijn maar één persoon bijna perfect geïdentificeerd kan worden, is voor die ene persoon op geen zinnige manier “99,9% privé”. En als die ene persoon voorspelbaar de patiënt met een zeldzame ziekte of uit een etnische minderheid is, is de maat niet alleen onvolledig maar stilletjes discriminerend. Hij rapporteert veiligheid voor de meerderheid en noemt dat veiligheid voor iedereen.

Dat is de verschuiving die het artikel afdwingt: van *hoe privé is dit model gemiddeld?* naar *wie is de meest blootgestelde patiënt, en wie is dat?* Dat zijn verschillende vragen, en alleen de tweede is werkelijk een privacyvraag.

## Wat dit *niet* bewijst — of beweert

- Het zegt **niet** dat medische AI moet worden opgegeven. De auteurs praten over mitigatie, niet terugtrekking: meet en repareer het risico, stop niet met bouwen.

- Het betekent **niet** dat elk medisch model lekt of dat een bepaald ingezet model al is aangevallen. Het laat zien dat het *risico bestaat en ongelijk verdeeld is*, en dat standaard geaggregeerde maten het missen.
- Het betekent **niet** dat jouw dossier nu al blootligt. De aanval vereist specifieke voorwaarden — toegang tot het model, een kandidaatdossier en eigen infrastructuur — en is geen terloopse mogelijkheid.
- Het laat **niet** zien dat de *inhoud* van patiëntendossiers uitlekt. Wat lekt is membership, het feit van deelname; dat is om een andere reden gevaarlijk, niet omdat het volledige dossier wordt uitgegeven.
- Het reduceert de ongelijkheid **niet tot één headlinegetal**. Het sterke, reproduceerbare resultaat is het *patroon*: geaggregeerde maten onderschatten individueel risico en ondervertegenwoordigde patiënten dragen het grootste deel ervan, over datasets en honderden modellen.

## Hoe sterk is het bewijs?

Dit is een empirisch, methodologisch resultaat en een robuust resultaat. Het patroon — geaggregeerde privacymaten onderschatten systematisch het risico per patiënt, het resterende risico concentreert zich op ondervertegenwoordigde groepen en wordt erger bij grotere modelcapaciteit — hield stand over **zeven datasets met verschillende gegevenstypen** en veel modellen per dataset. Juist die breedte maakt het meetresultaat geloofwaardig in plaats van een artefact van één dataset.

Twee eerlijke kanttekeningen. Ten eerste is dit een demonstratie van *risico*, gemeten met aanvallen die de auteurs zelf uitvoeren; het karakteriseert hoe goed een capabele aanvaller *zou kunnen* presteren onder de genoemde aannames, niet hoe vaak zulke aanvallen in de echte wereld plaatsvinden. Ten tweede beschrijven de opvallende cijfers — bijna perfecte aanvalsprestaties voor sommige patiënten — bewust de slechtst beschermde personen. Dat is precies het punt. Ze moeten worden gelezen als “de staart is veel zwaarder dan het gemiddelde doet vermoeden”, niet als “de meeste patiënten zijn blootgesteld”.

De juiste houding is noch alarmisme noch wegwuiven: dit is een zorgvuldige studie over meerdere datasets die laat zien dat de standaardmanier waarop we privacy van medische AI certificeren blind is voor haar eigen slechtste gevallen — en dat die gevallen juist terechtkomen bij patiënten met de minste marge.

## Waarom het ertoe doet

Twee dingen maken dit meer dan een technische voetnoot.

Ten eerste verandert het wat “privacybeschermend” zou mogen betekenen. Als een model wordt uitgebracht met één gemiddelde privacyscore, kan die score werkelijk laag zijn en toch een subgroep verbergen die door een membership-aanval bijna perfect identificeerbaar is. De praktische eis van het artikel is concreet: beoordeel privacyrisico **per patiënt** vóór uitrol, beperk wie toegang heeft

tot ingezette modellen en gebruik technieken als **differential privacy** — kleine, zorgvuldig gekalibreerde ruis tijdens training die membership-aanvallen afzwakt, tegen een echte en beheerde prijs in modelnut. De afweging privacy–bruikbaarheid is expliciet en niet gratis. “We hebben het gemiddelde gecontroleerd” zou niet langer als controle mogen gelden.

Ten tweede vlecht het privacy samen met eerlijkheid. De groepen die in medische data ondervertegenwoordigd zijn — en daardoor al slechter door medische AI kunnen worden bediend — blijken ook degenen van wie die AI de privacy het minst goed beschermt. Een veld dat hard aan de eerste ongelijkheid werkt, kan de tweede niet bij iemand anders neerleggen. Het zijn dezelfde mensen.

Het geruststellende verhaal over medische-AI-privacy — *we hebben gemeten, het gemiddelde is laag, alles goed* — is precies het verhaal dat dit artikel uit elkaar haalt. Niet om mensen van de technologie weg te jagen, maar om de norm te verplaatsen naar waar ze hoort: een belofte die je alleen kunt zeggen te houden als je haar ook hebt gecontroleerd voor de persoon die het waarschijnlijkst schade ondervindt.

# Heldere samenvatting

Membership-inference-aanvallen proberen te bepalen of het dossier van een specifieke persoon in de trainingsdata van een AI-model zat. Omdat membership zelf iets kan verraden — bijvoorbeeld dat je een bepaalde ziekte had — is dat een echt privacylek, ook wanneer het onderliggende dossier nooit verschijnt. Eerder werk mat hoe vaak zulke aanvallen *gemiddeld* over een dataset slagen. Deze studie mat het *per patiënt*, over zeven medische datasets (beeldvorming, ECG, elektronische dossiers) en veel modellen per dataset, en vond dat geaggregeerde maten het risico sterk onderschatten: sommige individuen kunnen bijna perfect worden geïdentificeerd (aanval-AUC $\geq$ 0,95), zelfs wanneer het datasetgemiddelde veilig lijkt; de kwetsbaarsten komen systematisch uit ondervertegenwoordigde groepen (etnische minderheden, zeldzame ziekten, ongebruikelijke beeldkenmerken); en het probleem neemt toe met modelgrootte. De auteurs pleiten niet voor het opgeven van medische AI, maar voor privacy meten op individueel niveau, modeltoegang beheersen en differential privacy gebruiken. De kern: “gemiddeld privé” is geen privacygarantie, en juist de al minst beschermde patiënten vallen door het gat.

# Zonder omwegen

**Wat het artikel laat zien:** Over zeven medische datasets en veel modellen per dataset is het membership-inferentierisico voor sommige patiënten veel groter dan geaggregeerde maten suggereren — tot bijna perfecte aanvalsprestatie (AUC $\geq$ 0,95) — en dit resterende risico treft ondervertegenwoordigde groepen onevenredig en groeit met modelcapaciteit.

**Wat aannemelijk is maar niet bewezen:** Dat aanvallers dit al tegen ingezette klinische modellen gebruiken; de studie demonstreert de *mogelijkheid en verdeling* onder expliciete aannames, niet de frequentie van echte aanvallen.

**Wat het niet laat zien:** Dat medische AI moet worden opgegeven; dat elk model lekt of een specifiek ingezet model is gekraakt; dat de *inhoud* van dossiers in plaats van membership wordt blootgelegd; één universeel ongelijkheidsgetal — het robuuste resultaat is het patroon, niet één cijfer.

**Belangrijkste beperkingen:** Het meet worst-case-risico met de aanvallen van de auteurs zelf en geen geobserveerde incidenten; de dramatische cijfers beschrijven bewust de meest blootgestelde individuen; de exacte groepsverschillen worden als consistent patroon over datasets getoond in plaats van in één getal samengevat.

**Hoeveel vertrouwen mag een algemene lezer hierin hebben?** Hoog vertrouwen dat geaggregeerde privacymaten individueel risico onderschatten, dat het resterende risico zich concentreert op ondervertegenwoordigde patiënten en dat dit erger wordt bij grotere modellen — het patroon is aangetoond over veel datasets en modellen. Matig vertrouwen over hoe vaak zulke aanvallen werkelijk plaatsvinden, want dat meet deze studie niet. De veilige lezing is niet “medische AI lekt jouw data” en ook niet “privacy is opgelost”, maar: “de standaardprivacycheck is blind voor zijn eigen slechtste gevallen, en die gevallen vallen op de kwetsbaarste patiënten — dus de check moet veranderen.”

---

## Bronnen

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