The Clean Paper

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

En medicinsk AI kan beskytte privatlivet i gennemsnit og stadig blotlægge enkelte patienter — især de underrepræsenterede

At en medicinsk AI-model kaldes »privatlivsbevarende«, hviler som regel på ét gennemsnitstal. Dette studie hævder, at det er den forkerte test. Forfatterne undersøgte medlemskabsinferensangreb — som afslører, om en bestemt persons oplysninger var med i en models træningsdata, og dermed kan røbe, at personen havde en bestemt sygdom — og målte risikoen for hver patient frem for samlet, på tværs af syv medicinske datasæt (billeddiagnostik, EKG, elektroniske journaler) og mange modeller pr. datasæt. Mønstret: Modeller, der ser sikre ud i gennemsnit, kan stadig lade en angriber identificere bestemte personer næsten perfekt (angrebs-AUC $\geq 0{,}95$); de eksponerede kommer systematisk fra underrepræsenterede grupper (etniske minoriteter, sjælden sygdom, usædvanlig billeddiagnostik); og problemet forværres, når modellerne vokser. Forfatterne siger ikke, at medicinsk AI skal opgives — de siger, at privatlivsrisikoen skal måles for hver patient, at adgangen til modeller skal kontrolleres, og at differentiel privatlivsbeskyttelse skal bruges. Den ubehagelige kerne: »privat i gennemsnit« er ingen privatlivsgaranti, og det svigter de patienter, der allerede er dårligst beskyttet.

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

En privatlivsgaranti er et løfte til en person, ikke til et gennemsnit

Når en medicinsk AI-model kaldes »privatlivsbevarende«, hviler påstanden som regel på ét tal: Blandt alle de patienter, hvis data blev brugt til at træne modellen, er den gennemsnitlige sandsynlighed lav for, at en enkelt persons medlemskab kan udledes. Det lyder betryggende. Studiets stille, ubehagelige pointe er, at det er det forkerte tal.

Privatliv er ikke et gennemsnit. Det er et løfte til hvert enkelt menneske — at det at være med i træningsdatasættet ikke senere vil føre til, at personen bliver afsløret. Og et gennemsnit kan holde dette løfte for næsten alle, samtidig med at det brydes fuldstændigt for nogle få. Forskerne målte privatlivsrisikoen én patient ad gangen, med en detaljeringsgrad ingen havde brugt før, og fandt netop dette: Modeller, der ser sikre ud samlet set, kan lække bestemte personers medlemskab næsten perfekt — og de personer, der afsløres, er uforholdsmæssigt ofte dem, der allerede er dårligst beskyttet.

Hvad forfatterne gjorde

Holdet — ledet af Moritz Knolle, med Daniel Rückert (begge ved Münchens Tekniske Universitet) og Georg Kaissis (Hasso Plattner-instituttet), sammen med kolleger ved Imperial College London — tog en velkendt trussel mod privatlivet, et såkaldt **medlemskabsinferensangreb**, og ændrede spørgsmålet. I stedet for at spørge »hvor ofte lykkes dette angreb i gennemsnit på tværs af datasættet?« spurgte de »hvor udsat er *netop denne patient*?«

De gennemførte denne analyse for hver enkelt patient i **syv etablerede medicinske datasæt**, der omfattede meget forskellige datatyper — medicinsk billeddiagnostik, elektrokardiogrammer og elektroniske patientjournaler — og 200 modeller pr. datasæt. For hver patient i hver model estimerede de, hvor sikkert en angriber kunne afgøre, om personens oplysninger havde været en del af træningsdataene. Derefter opdelte de resultaterne efter gruppe: sygdomsstatus, selvrapporteret race, forsikring, køn og billeddannelsesprotokol.

> **Hvad et medlemskabsinferensangreb faktisk er**
>
> Lækagen her er mere subtil end, at »modellen spytter din journal ud«. Den handler om *medlemskab* — selve det faktum, at dine data fandtes i træningsdatasættet.
> Begynd med, hvad en af disse modeller normalt gør. Du giver den et billede — for eksempel et røntgenbillede af brystkassen — og den giver sandsynligheder tilbage: 78 % sandsynlighed for lungebetændelse, sammen med vurderinger af kardiomegali, ødem og konsolidering. Angrebet bruger *kun* dette helt almindelige diagnostiske svar. Intet eksotisk.
> Den svaghed, angrebet udnytter, er denne: En model er som regel lidt **mere sikker** på præcis de eksempler, den blev trænet på, end på eksempler den aldrig har set. Tænk på en studerende, der i hemmelighed har set eksamen på forhånd — på de spørgsmål, den studerende allerede havde øvet sig på, kommer svarene lidt for hurtigt og lidt for selvsikkert. At finde ud af ud

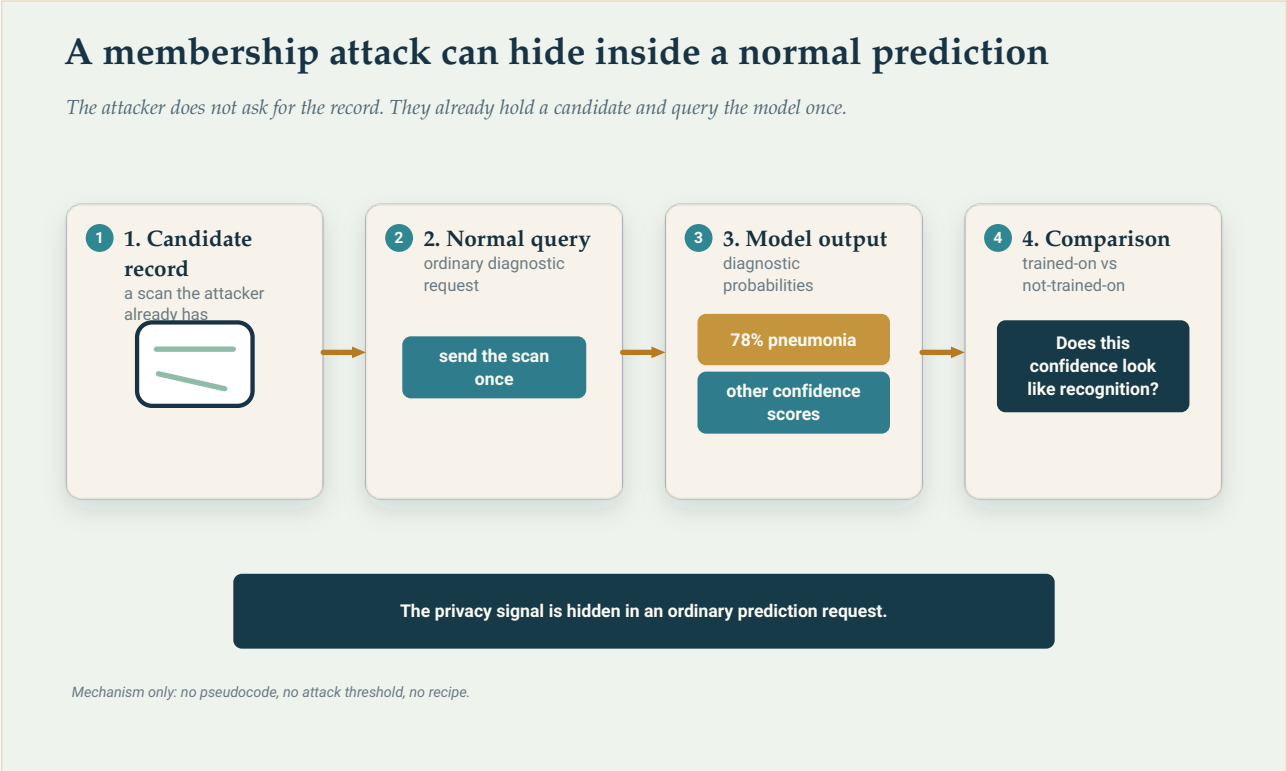
fra dette tegn, om en bestemt oplysning var blandt de eksempler, modellen studerede, er det, »medlemskabsinferens« betyder.

Sådan foregår angrebet trin for trin. Nogen vil vide, om en bestemt persons billede blev brugt til at træne modellen. De beder **ikke** modellen om oplysningerne — de har allerede et muligt billede (personens eget eller en tæt kopi). De sender det ind én gang som en almindelig diagnostisk forespørgsel og noterer, hvor sikkert svaret er. Derefter undersøger de, om denne sikkerhed ligner en model, der *havde* trænet på billedet, mere end en, der *ikke havde* — en sammenligning, de billigt kan foretage ved at træne deres egen stedfortræder (en »referencemodel«) til at lære, hvordan denne afslørende oversikkerhed ser ud, på en almindelig computer uden særlig hardware. Hvis den virkelige model er usædvanligt sikker på netop dette billede — som om den genkender det — er det stærk evidens for, at billedet var med i træningsdataene.

Det urovækkende er, at intet ved forespørgslen ligner et angreb: Det er den samme diagnostiske forespørgsel, som en kliniker ville sende, og privatlivssignalet er skjult i en almindelig forudsigelse. Netop derfor er risikoen konkret — selv om den hidtil er blevet påvist under angivne laboratorieantagelser og ikke opdaget ude i virkeligheden.

Hvorfor betyder medlemskab noget, hvis selve journaloplysningen aldrig lækker? Fordi *medlemskab er en kendsgerning*. Hvis en model blev trænet på patienter, der fik en bestemt immunterapi mod kræft, afslører en bekræftelse af, at dine oplysninger er med, at du sandsynligvis havde denne kræftsygdom — noget, et forsikringsselskab eller en arbejdsgiver aldrig burde kunne udlede. Indholdet forbliver forseglet; det er tilhørsforholdet, der slipper ud.

Forfatterne fremlægger disse antagelser tydeligt — adgang til modellens almindelige forudsigelser, en mulig journaloplysning og angriberens egen referencemodel — ikke som en vejledning, men for at truslen kan vurderes sagligt i stedet for at blive affejet med en håndbevægelse.



Angrebet kræver ikke et særligt privatlivs-API. En almindelig diagnostisk forespørgsel kan lække et medlemskabssignal, hvis modellen er usædvanligt sikker på en oplysning, den har set før.

Hvad de fandt

Tre fund, hvert skarpere end det foregående.

Gennemsnit skjuler de udsatte. Målt samlet — den gængse metode — ser mange af disse modeller betryggende private ud: For de fleste patienter klarer angrebet sig ikke bedre end tilfældigheder. Billedet for hver enkelt patient var et andet. Som Knolle udtrykte det i annonceringen af studiet, har tidligere vurderinger »kun målt den gennemsnitlige risiko på tværs af alle patienter. Vi undersøgte for første gang risikoen på den enkelte patients niveau — og det tegner et helt andet billede.« For nogle personer lykkes angrebet **næsten perfekt** — forfatterne måler det som en angrebs-AUC på **0,95 eller højere** (0,5 svarer til plat eller krone, 1,0 er fejlfrit) — selv når gennemsnittet for hele datasættet ikke så bedre ud end tilfældigheder.

Gennemsnittet kan ligge nær tilfældighedsniveauet, samtidig med at en lille hale af oplysninger er langt mere udsat. Studiets pointe er, at privatlivsbeskyttelsen skal undersøges på patientniveau, ikke kun samlet.

Eksponeringen er ulige — og den rammer de underrepræsenterede. De patienter, der var mest sårbare over for et næsten perfekt angreb, tilhørte systematisk grupper, som var **underrepræsenteret i dataene**: etniske minoriteter, fænotyper ved sjældne sygdomme og usædvanlige billedkarakteristika. I datasættet med elektroniske journaler forekom sorte patienter **31 % oftere** end forventet blandt de mest sårbare oplysninger; i mammografidatasættet var billeder, der var markeret som mistænkelige for malignitet, overrepræsenteret i denne farezone med hele **+1.179 %**. (Disse to tal er eksempler, ikke hele billedet — studiet rapporterer den samme skævhed i flere grupper — og de er *relative* overrepræsentationer inden for den lille hale med ekstrem risiko: hvor meget oftere disse patienter forekommer blandt de mest eksponerede oplysninger, ikke andelen af sorte patienter eller patienter med mistænkelige mammografibilleder, som er eksponeret.) En model har færre lignende eksempler, som disse patienter kan falde i ét med, så deres oplysninger skiller sig ud — og det at skille sig ud er netop, hvad et medlemskabsangreb opdager. Svigtet i privatlivsbeskyttelsen er ikke tilfældigt; det samler sig hos de mennesker, der allerede befinder sig i udkanten.

Større modeller gør det værre. Antallet af patienter, der var udsat for næsten perfekte angreb, steg kraftigt med modellens kapacitet — i ét dermatologisk datasæt voksede andelen fra stort set nul i den mindste model til omkring **én ud af ti** i den største. Efterhånden som medicinske AI-modeller bliver større og mere kapable — den retning hele feltet bevæger sig i — bliver denne specifikke risiko, advarer forfatterne, mere alvorlig, ikke mindre.

Rückerts sammenfatning er kontant: »Det er ikke en acceptabel risiko. Sundhedsdata er yderst følsomme.«

Hvorfor »sikker i gennemsnit« er den forkerte test

Det er fristende at læse en lav gennemsnitlig risiko som en blåstempling. Studiets centrale lære er, at et gennemsnit her ikke blot er upræcist — det måler det forkerte.

En privatlivsgaranti er kun meningsfuld, hvis den gælder for den person, der er mest udsat, ikke personen i midten. En model, hvor 999 ud af 1.000 patienter ikke kan identificeres, men én kan identificeres næsten perfekt, er ikke »99,9 % privat« i nogen betydning, der har værdi for den ene person — og hvis denne person forudsigeligt er patienten med en sjælden sygdom eller patienten fra en etnisk minoritet, er målet ikke bare ufuldstændigt, men stiltiende diskriminerende. Det rapporterer sikkerhed for flertallet og kalder det sikkerhed for alle.

Det er det skift, studiet fremtvinger: fra *hvor privat er denne model i gennemsnit?* til *hvilken patient er mest udsat, og hvem er personen?* Det er forskellige spørgsmål, og kun det andet er et spørgsmål om privatliv.

Hvad dette *ikke* beviser — eller hævder

- Det siger **ikke**, at medicinsk AI bør opgives. Forfatternes ramme er risikobegrænsning, ikke tilbagetog: Mål og afhjælp risikoen, hold ikke op med at bygge.
- Det betyder **ikke**, at alle medicinske modeller lækker, eller at en bestemt ibrugtaget model er blevet angrebet. Det viser, at *risikoen findes og er ulige fordelt*, og at almindelige aggregerede mål overser den.
- Det betyder **ikke**, at dine journaloplysninger allerede er eksponeret. Angrebet kræver bestemte betingelser — adgang til modellen, en mulig journaloplysning og angriberens egen infrastruktur — og er ikke noget, man uden videre kan gøre.
- Det viser **ikke**, at patienternes *indhold* lækker. Det, der lækker, er medlemskab — det faktum, at oplysningerne indgår — hvilket er farligt af en anden grund, ikke fordi journalen bliver udleveret.
- Det reducerer **ikke** forskellene til ét enkelt overskriftstal. Det stærke, reproducerbare resultat er *mønstret* — aggregerede mål undervurderer den individuelle risiko, og underrepræsenterede patienter bærer den største del af den — på tværs af datasæt og hundredvis af modeller.

Hvor stærk er evidensen?

Dette er et empirisk, metodologisk resultat, og det er robust. Mønstret — at aggregerede mål for privatlivsbeskyttelse systematisk undervurderer risikoen for den enkelte patient, at den resterende risiko samler sig i underrepræsenterede grupper, og at den forværres med modellens kapacitet — gik igen på tværs af **syv datasæt med forskellige datatyper** og et stort antal modeller pr. datasæt. Netop denne bredde gør måleresultatet troværdigt, frem for at det fremstår som et særtilfælde i ét datasæt.

To ærlige forbehold. For det første er dette en demonstration af *risiko*, målt ved at køre de angreb,

forfatterne selv byggede; den beskriver, hvor godt en kompetent angriber *kunne* klare sig under angivne antagelser, ikke hvor ofte virkelige angreb sker. For det andet beskriver de slående tal — næsten perfekt angrebssucces for nogle patienter — de mest udsatte personer *bevidst*; det er hele pointen, og tallene bør læses som »halen er langt tungere, end gennemsnittet antyder«, ikke som »de fleste patienter er eksponeret«.

Den passende holdning er hverken alarmisme eller afvisning: Dette er et omhyggeligt studie af flere datasæt, som viser, at standardmåden at certificere privatlivsbeskyttelse i medicinsk AI på er blind for sine egne værste tilfælde — og at disse værste tilfælde rammer de patienter, der har mindst spillerum.

Hvorfor det betyder noget

To ting gør dette til mere end en teknisk fodnote.

For det første ændrer det, hvad »privatlivsbevarende« bør have lov til at betyde. Hvis en model udgives med en gennemsnitlig privatlivsscore, kan denne score reelt være lav og stadig skjule en delmængde af patienter, som kan identificeres næsten perfekt med et medlemskabsangreb. Studiets praktiske krav er konkret: Vurder privatlivsrisikoen **for hver patient** før udgivelse, kontrollér, hvem der kan få adgang til ibrugtagne modeller, og brug teknikker som **differentiel privatlivsbeskyttelse** — små, omhyggeligt kalibrerede støjbidrag, der tilføjes under træningen og svækker medlemskabsangreb, med en reel og håndteret omkostning for modellens anvendelighed (afvejningen mellem privatliv og anvendelighed er udtrykkelig, ikke gratis). »Vi kontrollerede gennemsnittet« bør ikke længere tælle som at have kontrolleret.

For det andet fletter studiet privatliv sammen med retfærdighed. De samme grupper, der er underrepræsenteret i medicinske data — og derfor allerede betjenes dårligere af medicinsk AI — viser sig at være dem, hvis privatliv denne AI beskytter mindst. Et felt, der arbejder hårdt på den første ulighed, kan ikke behandle den anden som en andens ansvarsområde. Det er de samme mennesker.

Den betryggende fortælling om privatliv i medicinsk AI — *vi målte det, gennemsnittet er lavt, alt er fint* — er den fortælling, dette studie skiller ad. Ikke for at skræmme nogen væk fra teknologien, men for at flytte standarden derhen, hvor den hører hjemme: til et løfte, man kun kan hævde at holde, hvis man har kontrolleret det for den person, der har størst risiko for at blive skadet.

Kort fortalt

Medlemskabsinferensangreb forsøger at afgøre, om en bestemt persons oplysninger var med i en AI-models træningsdata — og fordi medlemskab i sig selv kan afsløre noget (for eksempel at personen havde en bestemt sygdom), er det et reelt brud på privatlivet, selv når den underliggende journaloplysning aldrig kommer frem. Tidligere forskning målte, hvor ofte sådanne angreb lykkes *i gennemsnit* på tværs af et datasæt. Dette studie målte det *for hver patient* på tværs af syv medicinske datasæt (billeddiagnostik, EKG, elektroniske journaler) og mange modeller pr. datasæt og fandt, at

aggregerede mål kraftigt undervurderer risikoen: Nogle personer kan identificeres næsten perfekt (angrebs-AUC $\geq$ 0,95), selv når gennemsnittet for hele datasættet ser sikkert ud; de mest sårbare kommer systematisk fra underrepræsenterede grupper (etniske minoriteter, sjælden sygdom, usædvanlig billeddiagnostik); og problemet vokser med modellens størrelse. Forfatterne argumenterer ikke for at opgive medicinsk AI — de argumenterer for at måle privatlivsrisiko på individniveau, kontrollere adgangen til modeller og bruge differentiel privatlivsbeskyttelse. Konklusionen: »privat i gennemsnit« er ingen privatlivsgaranti, og de, den svigter, er som regel dem, der allerede er dårligst beskyttet.

Uden omsvøb

Hvad studiet viser: På tværs af syv medicinske datasæt og mange modeller pr. datasæt er risikoen for medlemskabsinferens for den enkelte patient langt højere for nogle personer, end aggregerede mål antyder — op til næsten perfekt angrebssucces (AUC $\geq$ 0,95) — og denne resterende risiko rammer underrepræsenterede grupper uforholdsmæssigt og vokser med modellens kapacitet.

Hvad der er plausibelt, men ikke bevist: At virkelige angribere allerede udnytter dette mod ibrugtagne kliniske modeller; studiet viser *kapaciteten og dens fordeling* under angivne antagelser, ikke hvor ofte angreb sker i virkeligheden.

Hvad det ikke viser: At medicinsk AI bør opgives; at alle modeller lækker, eller at en bestemt ibrugtaget model har været udsat for et brud; at *indholdet* i patientjournaler (frem for medlemskabet) eksponeres; ét enkelt tal for forskellen — det robuste resultat er mønstret, ikke ét tal.

Vigtigste begrænsninger: Det måler værst tænkelige risiko med forfatternes egne angreb, ikke observerede hændelser; de dramatiske tal beskriver bevidst de mest eksponerede personer; og de præcise forskelle mellem grupper vises som et konsekvent mønster på tværs af datasæt frem for at blive reduceret til ét tal.

Hvor stor tillid bør en almindelig læser have? Høj tillid til, at aggregerede mål for privatlivsbeskyttelse undervurderer den individuelle risiko, at den resterende risiko samler sig hos underrepræsenterede patienter, og at dette bliver værre med modellens størrelse — det er påvist på tværs af mange datasæt og modeller. Moderat tillid til, hvor ofte sådanne angreb sker i virkeligheden, hvilket dette studie ikke måler. Den sikre læsning er ikke »medicinsk AI lækker dine data« og ikke »privatlivsproblemet er løst«, men »standardkontrollen af privatlivsbeskyttelse er blind for sine egne værste tilfælde, og disse tilfælde rammer de mest sårbare patienter — derfor skal kontrollen ændres.«

Kilder

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