The Clean Paper

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

En medisinsk KI kan ivareta personvernet i gjennomsnitt og likevel eksponere enkeltpasienter — særlig de underrepresenterte

At en medisinsk KI-modell kalles «personvernbevarende», hviler vanligvis på ett gjennomsnittstall. Denne studien hevder at dette er feil test. Forfatterne studerte medlemskapsinferensangrep — som avslører om opplysningene til en bestemt person var med i treningsdataene til en modell og dermed kan røpe at personen hadde en bestemt sykdom — og målte risikoen for hver enkelt pasient i stedet for samlet, på tvers av sju medisinske datasett (bildediagnostikk, EKG, elektroniske journaler) og mange modeller per datasett. Mønsteret: Modeller som ser trygge ut i gjennomsnitt, kan likevel la en angriper identifisere bestemte personer nesten perfekt (angreps-AUC ≥ 0,95); de eksponerte kommer systematisk fra underrepresenterte grupper (etniske minoriteter, sjelden sykdom, uvanlig bildediagnostikk); og problemet blir verre når modellene vokser. Forfatterne sier ikke at medisinsk KI skal forlates — de sier at personvern må måles for hver enkelt pasient, at tilgangen til modeller må kontrolleres, og at differensielt personvern må brukes. Den ubehagelige kjernen: «privat i gjennomsnitt» er ingen personverngaranti, og den svikter pasientene som allerede har minst beskyttelse.

Written by Lucio Vaglio · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso · 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 personverngaranti er et løfte til en person, ikke til et gjennomsnitt

Når en medisinsk KI-modell kalles «personvernbevarende», hviler påstanden vanligvis på ett tall: Blant alle pasientene hvis data ble brukt til å trene modellen, er den gjennomsnittlige sannsynligheten lav for at medlemskapet til en enkeltperson kan utledes. Det høres betryggende ut. Studiens stillferdige, ubehagelige poeng er at dette er feil tall.

Personvern er ikke et gjennomsnitt. Det er et løfte til hver enkelt — at det å være med i treningsdatasettet ikke senere skal føre til at personen blir avslørt. Og et gjennomsnitt kan holde dette løftet for nesten alle, samtidig som det brytes fullstendig for noen få. Forskerne målte personvernrisikoen én pasient om gangen, med en detaljgrad ingen hadde brukt før, og fant nettopp dette: Modeller som ser trygge ut samlet sett, kan lekke medlemskapet til bestemte personer nesten perfekt — og personene som røpes, er i uforholdsmessig stor grad dem som allerede har minst beskyttelse.

Hva forfatterne gjorde

Teamet — ledet av Moritz Knolle, med Daniel Rückert (begge ved Münchens tekniske universitet) og Georg Kaissis (Hasso Plattner-instituttet), sammen med kolleger ved Imperial College London — tok for seg en velkjent personverntrussel som kalles et **medlemskapsinferensangrep**, og endret spørsmålet. I stedet for å spørre «hvor ofte lykkes dette angrepet i gjennomsnitt i hele datasettet?», spurte de «hvor utsatt er *akkurat denne pasienten*?»

De gjennomførte denne analysen for hver enkelt pasient i **sju etablerte medisinske datasett**, som omfattet svært forskjellige datatyper — medisinsk bildediagnostikk, elektrokardiogrammer og elektroniske pasientjournaler — og 200 modeller per datasett. For hver pasient i hver modell anslo de hvor sikkert en angriper kunne avgjøre om opplysningene til personen hadde vært en del av treningsdataene. Deretter fordelte de resultatene på grupper: sykdomsstatus, selvrapportert rase, forsikring, kjønn og bildeopptaksprotokoll.

Hva et medlemskapsinferensangrep faktisk er

Lekkasjen her er mer subtil enn at «modellen spytter ut journalen din». Den handler om *medlemskap* — selve det faktum at dataene dine var med i treningsdatasettet.
Begynn med hva en slik modell vanligvis gjør. Du gir den et bilde — for eksempel et røntgenbilde av brystkassen — og får sannsynligheter tilbake: 78 % sannsynlighet for lungebetennelse, sammen med vurderinger av kardiomegali, ødem og konsolidering. Angrepet bruker *bare* dette alminnelige diagnostiske svaret. Ikke noe eksotisk.
Svakheten det utnytter, er denne: En modell er vanligvis litt **sikrere** på akkurat de eksemplene den ble trent på enn på eksempler den aldri har sett. Tenk på en student som i hemmelighet fikk se eksamen på forhånd — på spørsmålene studenten allerede hadde øvd på, kommer svarene litt for raskt og litt for selvsikkert. Å finne ut fra dette kjennetegnet om en bestemt

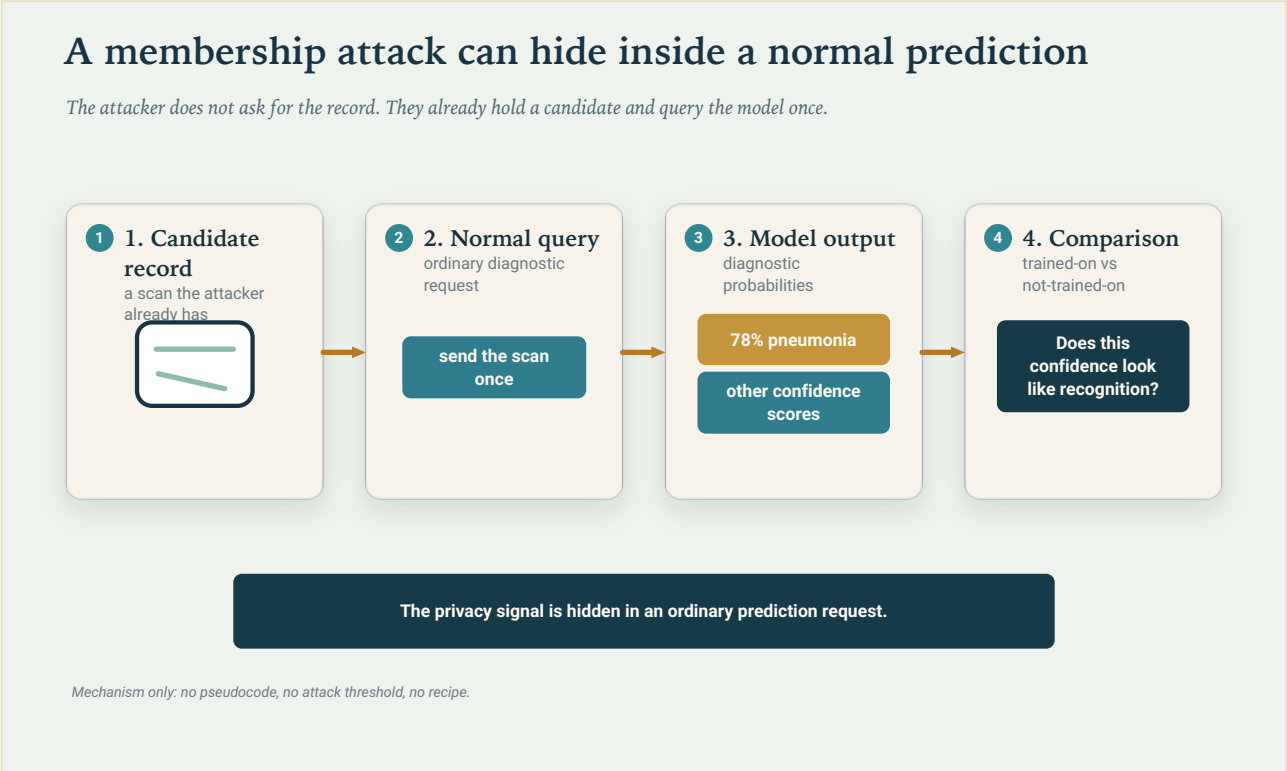
oppføring var blant eksemplene modellen studerte, er det «medlemskapsinferens» betyr.

Slik foregår angrepet, trinn for trinn. Noen ønsker å vite om bildet til en bestemt person ble brukt til å trene modellen. De ber **ikke** modellen om å gi dem opplysningene — de har allerede et aktuelt bilde (personens eget eller en nær kopi). De sender det inn én gang som en vanlig diagnostisk forespørsel og merker seg hvor sikkert svaret er. Deretter undersøker de om denne sikkerheten ligner mer på en modell som *hadde* trent på bildet, eller en som *ikke hadde* gjort det — en sammenligning de kan gjøre billig ved å trene sin egen stedfortreder (en «referansemodell») til å lære hvordan denne avslørende oversikkerheten ser ut, på en vanlig datamaskin uten spesialutstyr. Hvis den virkelige modellen er uvanlig sikker på akkurat dette bildet — som om den kjenner det igjen — er det sterke holdepunkter for at bildet var med i treningsdataene.

Det urovekkende er at ingenting ved forespørselen ser ut som et angrep: Det er den samme diagnostiske forespørselen som en kliniker ville ha sendt, og personvernsignalet er skjult i en vanlig prediksjon. Nettopp derfor er risikoen konkret — selv om den hittil er påvist under oppgitte laboratorieforutsetninger og ikke oppdaget ute i virkeligheten.

Hvorfor betyr medlemskap noe hvis selve journalopplysningen aldri lekker? Fordi *medlemskap er et faktum*. Hvis en modell ble trent på pasienter som fikk en bestemt immunterapi mot kreft, vil en bekreftelse på at opplysningene dine er med, avsløre at du sannsynligvis hadde denne kreftsykdommen — noe et forsikringsselskap eller en arbeidsgiver aldri burde kunne utlede. Innholdet forblir forseglet; det er tilhørigheten som slipper ut.

Forfatterne gjør tydelig rede for disse forutsetningene — tilgang til modellens vanlige prediksjoner, en aktuell journaloppføring og angriperens egen referansemodell — ikke som en fremgangsmåte, men for at trusselen skal kunne vurderes saklig i stedet for å avfeies med en håndbevegelse.



Angrepet trenger ikke et eget personvern-API. En vanlig diagnostisk forespørsel kan lekke et medlemskapssignal hvis modellen er uvanlig sikker på en oppføring den har sett før.

Hva de fant

Tre funn, hvert skarpere enn det forrige.

Gjennomsnitt skjuler de utsatte. Målt samlet — den vanlige måten — ser mange av disse modellene betryggende private ut: For de fleste pasienter gjør angrepet det ikke bedre enn tilfeldigheter. Bildet for hver enkelt pasient var et annet. Som Knolle sa i kunngjøringen av studien, har tidligere vurderinger «bare målt den gjennomsnittlige risikoen på tvers av alle pasienter. Vi undersøkte for første gang risikoen på individnivå — og det tegner et helt annet bilde.» For enkelte personer lykkes angrepet **nesten perfekt** — forfatterne måler det som en angreps-AUC på **0,95 eller høyere** (0,5 tilsvarer myntkast, 1,0 er feilfritt) — selv når gjennomsnittet for hele datasettet ikke så bedre ut enn tilfeldigheter.

Gjennomsnittet kan ligge nær tilfeldighetsnivået samtidig som en liten hale av oppføringer er langt mer utsatt. Studiens poeng er at personvernet må undersøkes på pasientnivå, ikke bare samlet.

Eksponeringen er ulik — og den rammer de underrepresenterte. Pasientene som var mest sårbare for et nesten perfekt angrep, tilhørte systematisk grupper som var **underrepresentert i dataene**: etniske minoriteter, fenotyper ved sjeldne sykdommer og uvanlige bildeegenskaper. I datasettet med elektroniske journaler forekom svarte pasienter **31 % oftere** enn forventet blant de mest sårbare oppføringene; i mammografidatasettet var bilder som var flagget som mistenkelige for malignitet, overrepresentert i denne faresonen med hele **+1 179 %**. (Disse to tallene er eksempler, ikke hele bildet — studien rapporterer den samme skjevheten i flere grupper — og de er *relative* overrepresentasjoner innenfor den lille halen med ekstrem risiko: hvor mye oftere disse pasientene forekommer blant de mest eksponerte oppføringene, ikke andelen svarte pasienter eller pasienter med mistenkelige mammografibilder som er eksponert.) En modell har færre lignende eksempler som disse pasientene kan gli inn blant, så oppføringene deres skiller seg ut — og det å skille seg ut er nettopp det et medlemskapsangrep oppdager. Personvernsvikten er ikke tilfeldig; den samler seg hos mennesker som allerede befinner seg i utkanten.

Større modeller gjør det verre. Antallet pasienter som var utsatt for nesten perfekte angrep, steg kraftig med modellkapasiteten — i ett dermatologisk datasett økte andelen fra praktisk talt null i den minste modellen til rundt **én av ti** i den største. Etter hvert som medisinske KI-modeller blir større og mer kapable — den retningen hele feltet beveger seg i — blir denne bestemte risikoen, advarer forfatterne, mer alvorlig, ikke mindre.

Rückerts oppsummering er kontant: «Dette er ikke en akseptabel risiko. Helsedata er svært sensitive.»

Hvorfor «trygg i gjennomsnitt» er feil test

Det er fristende å lese en lav gjennomsnittlig risiko som en friskmelding. Studiens hovedlærdom er at gjennomsnittsberegning her ikke bare er upresist — det måler feil ting.

En personverngaranti er bare meningsfull hvis den gjelder for personen med størst risiko, ikke personen i midten. En modell der 999 av 1 000 pasienter ikke kan identifiseres, men én kan identifiseres nesten perfekt, er ikke «99,9 % privat» på noen måte som har betydning for den ene personen — og hvis denne personen forutsigbart er pasienten med en sjelden sykdom eller pasienten fra en etnisk minoritet, er målet ikke bare ufullstendig, men stilltiende diskriminerende. Det rapporterer trygghet for flertallet og kaller det trygghet for alle.

Det er dette skiftet studien tvinger frem: fra *hvor privat er denne modellen i gjennomsnitt?* til *hvilken pasient er mest utsatt, og hvem er denne personen?* Dette er forskjellige spørsmål, og bare det andre er et personvernspørsmål.

Hva dette *ikke* beviser — eller hevder

- Det sier **ikke** at medisinsk KI bør forlates. Forfatternes innramming er risikoreduserende tiltak, ikke retrett: Mål og rett opp risikoen, ikke slutt å bygge.
- Det betyr **ikke** at alle medisinske modeller lekker, eller at en bestemt modell i bruk er blitt angrepet. Det viser at *risikoen finnes og er ulikt fordelt,* og at vanlige aggregerte mål ikke fanger den opp.
- Det betyr **ikke** at journalopplysningene dine allerede er eksponert. Angrepet krever bestemte forhold — tilgang til modellen, en aktuell journaloppføring og angriperens egen infrastruktur — og er ikke noe hvem som helst kan gjøre uten videre.
- Det viser **ikke** at pasientenes *innhold* lekker. Det som lekker, er medlemskap — det faktum at opplysningen er med — noe som er farlig av en annen grunn, ikke fordi journalen dumpes ut.
- Det reduserer **ikke** forskjellene til ett enkelt overskriftstall. Det sterke, reproduserbare resultatet er *mønsteret* — aggregerte mål undervurderer den individuelle risikoen, og underrepresenterte pasienter bærer mesteparten av den — på tvers av datasett og hundrevis av modeller.

Hvor sterke er bevisene?

Dette er et empirisk, metodologisk resultat, og et robust sådant. Mønsteret — at aggregerte personvernmål systematisk undervurderer risikoen for hver enkelt pasient, at den gjenværende risikoen samler seg i underrepresenterte grupper, og at den blir verre med modellkapasiteten — gikk igjen på tvers av **sju datasett med forskjellige datatyper** og et stort antall modeller per datasett. Det er nettopp denne bredden som gjør måleresultatet troverdig, fremfor at det fremstår som et utslag i ett enkelt datasett.

To ærlige forbehold. For det første er dette en demonstrasjon av *risiko*, målt ved å kjøre angrepene forfatterne selv bygde; den beskriver hvor godt en kapabel angriper *kunne* lykkes under oppgitte forutsetninger, ikke hvor ofte virkelige angrep skjer. For det andre beskriver de slående tallene — nesten perfekt angrepssuksess for enkelte pasienter — de mest utsatte personene *med hensikt*; det er hele poenget, og de bør leses som «halen er langt tyngre enn gjennomsnittet tilsier», ikke som «de fleste pasientene er eksponert».

Den passende holdningen er verken alarmisme eller avvisning: Dette er en omhyggelig studie av flere datasett som viser at standardmåten å sertifisere personvern i medisinsk KI på er blind for sine egne verste tilfeller — og at disse verste tilfellene rammer pasientene med minst slingringsmonn.

Hvorfor det betyr noe

To ting gjør dette til mer enn en teknisk fotnote.

For det første endrer det hva «personvernbevarende» bør få lov til å bety. Hvis en modell lanseres med en gjennomsnittlig personvernscore, kan denne scoren faktisk være lav og likevel skjule en undergruppe av pasienter som er nesten perfekt identifiserbare med et medlemskapsangrep. Studiens praktiske krav er konkret: Vurder personvernrisikoen **for hver pasient** før lansering, kontroller hvem som har tilgang til modeller i bruk, og bruk teknikker som **differensielt personvern** — liten, nøye kalibrert støy som legges til under treningen og svekker medlemskapsangrep, med en reell og håndtert kostnad for modellens nytteverdi (avveiningen mellom personvern og nytteverdi er uttrykkelig, ikke gratis). «Vi kontrollerte gjennomsnittet» bør ikke lenger telle som at man har kontrollert.

For det andre fletter studien personvern sammen med rettferdighet. De samme gruppene som er underrepresentert i medisinske data — og derfor allerede får dårligere hjelp av medisinsk KI — viser seg å være dem KI-en beskytter personvernet dårligst for. Et fagfelt som arbeider hardt med den første skjevheten, kan ikke behandle den andre som noen andres ansvarsområde. Det er de samme menneskene.

Den betryggende fortellingen om personvern i medisinsk KI — *vi målte det, gjennomsnittet er lavt, alt er i orden* — er fortellingen denne studien plukker fra hverandre. Ikke for å skremme noen bort fra teknologien, men for å flytte standarden dit den hører hjemme: til et løfte man bare kan hevde å holde hvis man har kontrollert det for personen som mest sannsynlig blir skadet.

Ren oppsummering

Medlemskapsinferensangrep prøver å avgjøre om opplysningene til en bestemt person var med i treningsdataene til en KI-modell — og fordi medlemskap i seg selv kan avsløre noe (for eksempel at personen hadde en bestemt sykdom), er dette en reell personvernlekkasje selv om den underliggende journalopplysningen aldri kommer frem. Tidligere forskning målte hvor ofte slike angrep lykkes *i gjennomsnitt* i et datasett. Denne studien målte det *for hver enkelt pasient*, på tvers av sju

medisinske datasett (bildediagnostikk, EKG, elektroniske journaler) og mange modeller per datasett, og fant at aggregerte mål kraftig undervurderer risikoen: Noen personer kan identifiseres nesten perfekt (angreps-AUC $\geq 0{,}95$) selv når gjennomsnittet for hele datasettet ser trygt ut; de mest sårbare kommer systematisk fra underrepresenterte grupper (etniske minoriteter, sjelden sykdom, uvanlig bildediagnostikk); og problemet vokser med modellstørrelsen. Forfatterne argumenterer ikke for å forlate medisinsk KI — de argumenterer for å måle personvern på individnivå, kontrollere tilgangen til modeller og bruke differensielt personvern. Konklusjonen: «privat i gjennomsnitt» er ingen personverngaranti, og de den svikter, er vanligvis dem som allerede har minst beskyttelse.

Uten omsvøp

Hva studien viser: På tvers av sju medisinske datasett og mange modeller per datasett er risikoen for medlemskapsinferens per pasient langt høyere for enkelte personer enn aggregerte mål antyder — opp til nesten perfekt angrepssuksess (AUC $\geq 0{,}95$) — og denne gjenværende risikoen rammer underrepresenterte grupper uforholdsmessig og vokser med modellkapasiteten.

Hva som er plausibelt, men ikke bevist: At virkelige angripere allerede utnytter dette mot kliniske modeller i bruk; studien viser *kapasiteten og hvordan den er fordelt* under oppgitte forutsetninger, ikke hvor ofte angrep skjer i virkeligheten.

Hva den ikke viser: At medisinsk KI bør forlates; at alle modeller lekker, eller at en bestemt modell i bruk har hatt et sikkerhetsbrudd; at *innholdet* i pasientjournaler (i stedet for medlemskapet) eksponeres; én enkelt tallfestet forskjell — det robuste resultatet er mønsteret, ikke ett tall.

Viktigste begrensninger: Den måler verstefallsrisiko med forfatternes egne angrep, ikke observerte hendelser; de dramatiske tallene beskriver med hensikt de mest eksponerte personene; og de nøyaktige forskjellene mellom grupper vises som et konsekvent mønster på tvers av datasett, i stedet for å reduseres til ett tall.

Hvor stor tillit bør en vanlig leser ha? Høy tillit til at aggregerte personvernmål undervurderer den individuelle risikoen, at den gjenværende risikoen samler seg hos underrepresenterte pasienter, og at dette blir verre med modellstørrelsen — det er påvist i mange datasett og modeller. Moderat tillit når det gjelder hvor ofte slike angrep skjer i virkeligheten, noe denne studien ikke måler. Den trygge lesningen er ikke «medisinsk KI lekker dataene dine» og ikke «personvernet er løst», men «standardkontrollen av personvern er blind for sine egne verste tilfeller, og disse tilfellene rammer de mest sårbare pasientene — derfor må kontrollen endres.»

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-ai

Medical Xpress (2026) — coverage of the study
https://medicalxpress.com/news/2026-06-patient-groups-vulnerable-privacy-medical.html