The Clean Paper

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

En medicinsk AI kan skydda integriteten i genomsnitt och ändå röja enskilda patienter — särskilt de underrepresenterade

Att en medicinsk AI-modell kallas ”integritetsbevarande” vilar vanligtvis på ett enda genomsnittstal. Den här studien hävdar att det är fel test. Författarna studerade medlemskapsinferensattacker — som avslöjar om en viss persons uppgifter ingick i en modells träningsdata och därmed kan röja att personen hade en viss sjukdom — och mätte risken per patient i stället för sammantaget, i sju medicinska datamängder (bilddiagnostik, EKG, elektroniska journaler) och många modeller per datamängd. Mönstret: modeller som ser säkra ut i genomsnitt kan ändå låta en angripare identifiera vissa individer nästan perfekt (attack-AUC ≥ 0,95); de utsatta kommer systematiskt från underrepresenterade grupper (etniska minoriteter, sällsynt sjukdom, ovanlig bilddiagnostik); och problemet förvärras när modellerna växer. Författarna säger inte att medicinsk AI ska överges — de säger att integriteten ska mätas per patient, att tillgången till modeller ska kontrolleras och att differentiell integritet ska användas. Den obekväma kärnan: ”privat i genomsnitt” är ingen integritetsgaranti, och det sviker de patienter som redan har sämst skydd.

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 integritetsgaranti är ett löfte till en person, inte till ett genomsnitt

När en medicinsk AI-modell kallas ”integritetsbevarande” vilar det påståendet vanligtvis på ett enda tal: bland alla patienter vars data användes för att träna modellen är den genomsnittliga sannolikheten för att någon enskild individs medlemskap i träningsdata kan härledas låg. Det låter betryggande. Studiens lågmälda, obekväma poäng är att det är fel tal.

Integritet är inget genomsnitt. Det är ett löfte till varje individ — att deras medverkan i träningsdata inte ska komma tillbaka och röja dem. Och ett genomsnitt kan hålla det löftet för nästan alla samtidigt som det bryts fullständigt för ett fåtal. Forskarna mätte integritetsrisken för en patient i taget, med en detaljnivå som ingen hade använt tidigare, och fann just detta: modeller som ser säkra ut sammantaget kan röja specifika individers medlemskap nästan perfekt — och de individer som röjs är oproportionerligt ofta de som redan har sämst skydd.

Vad författarna gjorde

Teamet — lett av Moritz Knolle, tillsammans med Daniel Rückert (båda vid Münchens tekniska universitet) och Georg Kaissis (Hasso Plattner-institutet), samt kollegor vid Imperial College London — tog ett välkänt integritetshot som kallas en **medlemskapsinferensattack** och ändrade frågan. I stället för att fråga ”hur ofta lyckas attacken i genomsnitt i hela datamängden?” frågade de ”hur utsatt är *just den här patienten*?”

De genomförde denna analys patient för patient i **sju etablerade medicinska datamängder**, med vitt skilda datatyper — medicinsk bilddiagnostik, elektrokardiogram och elektroniska patientjournaler — och i 200 modeller per datamängd. För varje patient i varje modell uppskattade de hur säkert en angripare kunde avgöra om personens uppgifter hade ingått i träningsdata. Därefter delade de upp resultaten efter grupp: sjukdomsstatus, självrapporterad ras, försäkring, kön och bildtagningsprotokoll.

Vad en medlemskapsinferensattack faktiskt är

Läckan här är mer subtil än att ”modellen spottar ur sig din journal”. Den gäller *medlemskap* — själva faktumet att dina data fanns i träningsuppsättningen.

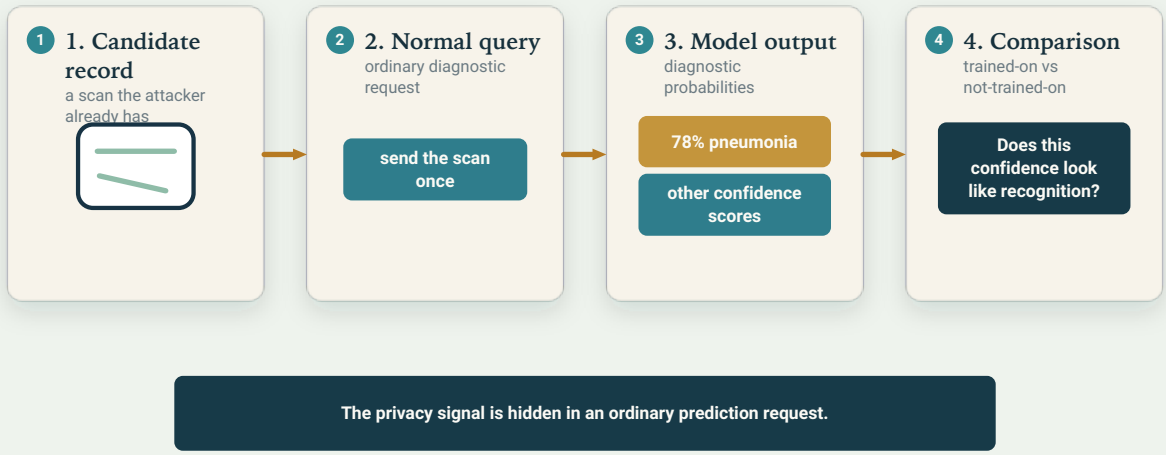
Börja med vad en sådan modell normalt gör. Du ger den en bild — till exempel en lungröntgen — och får tillbaka sannolikheter: 78 % sannolikhet för lunginflammation, tillsammans med bedömningar av kardiomegali, ödem och konsolidering. Attacken använder *enbart* detta vardagliga diagnostiska svar. Inget exotiskt.

Svagheten den utnyttjar är denna: en modell är vanligtvis lite **säkrare** på exakt de exempel som den tränades på än på sådana den aldrig har sett. Tänk på en student som i hemlighet sett tentamen i förväg — på frågorna som studenten redan har övat på kommer svaren lite för snabbt och lite för självsäkert. Att utifrån den ledtråden ta reda på om en viss uppgift var ett

av exemplen som modellen studerade är vad ”medlemskapsinferens” betyder.
Så här går attacken till, steg för steg. Någon vill veta om en viss persons bild användes för att träna modellen. Personen ber **inte** modellen om journaluppgiften — angriparen har redan en tänkbar bild (personens egen eller en nära kopia). Bilden skickas in en enda gång, som en vanlig diagnostisk förfrågan, och angriparen noterar hur säkert svaret är. Sedan kontrollerar angriparen om denna säkerhet liknar den hos en modell som *har* tränats på bilden eller en som *inte har* det — en jämförelse som går att göra billigt genom att träna en egen ersättningsmodell (en ”referensmodell”) för att lära sig hur denna avslöjande översäkerhet ser ut, på en vanlig dator utan särskild hårdvara. Om den verkliga modellen är ovanligt säker på just denna bild — som om den känner igen den — är det starka belägg för att bilden ingick i träningsdata.
Det oroande är att inget i förfrågan ser ut som en attack: det är samma diagnostiska fråga som en kliniker skulle ställa, och integritetssignalen är dold i en vanlig förutsägelse. Det är just därför risken är konkret — även om den hittills har visats under angivna laboratorieantaganden och inte upptäckts i verkligheten.
Varför spelar medlemskapet någon roll om själva journaluppgiften aldrig läcker? Därför att *medlemskapet är ett faktum*. Om en modell tränades på patienter som fick en viss immunterapi mot cancer, avslöjar en bekräftelse på att din uppgift finns med att du troligen hade den cancersjukdomen — något som ett försäkringsbolag eller en arbetsgivare aldrig borde kunna härleda. Innehållet förblir förseglat; det är tillhörigheten som avslöjas.
Författarna redovisar dessa antaganden tydligt — tillgång till modellens vanliga förutsägelser, en tänkbar journaluppgift och angriparens egen referensmodell — inte som en instruktion, utan för att hotet ska kunna analyseras i stället för att avfärdas med en handviftning.

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.



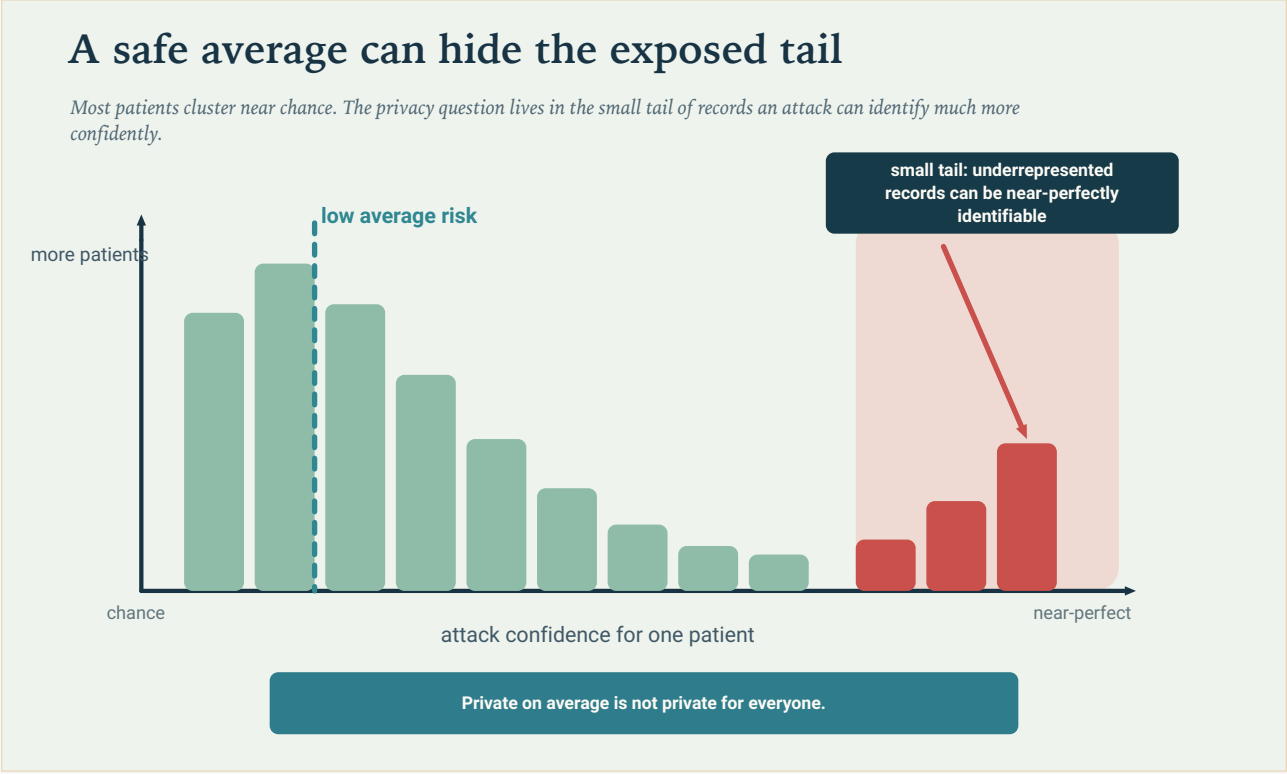
Mechanism only: no pseudocode, no attack threshold, no recipe.

Attacken behöver inget särskilt integritets-API. En vanlig diagnostisk förfrågan kan läcka en medlemskapssignal om modellen är ovanligt säker på en uppgift som den har sett förut.

Vad de fann

Tre fynd, vart och ett skarpare än det föregående.

Genomsnitt döljer de utsatta. Mätt sammantaget — det vanliga sättet — ser många av modellerna betryggande privata ut: för de flesta patienter lyckas attacken inte bättre än slumpen. Bilden patient för patient var en annan. Som Knolle uttryckte det i tillkännagivandet av studien har tidigare bedömningar ”bara någonsin mätt den genomsnittliga risken för alla patienter. Vi undersökte för första gången risken på individnivå — och det ger en helt annan bild.” För vissa individer lyckas attacken **nästan perfekt** — författarna mäter detta som en attack-AUC på **0,95 eller högre** (0,5 motsvarar slantsingling, 1,0 är felfritt) — trots att genomsnittet för hela datamängden inte såg bättre ut än slumpen.



Genomsnittet kan ligga nära slumpnivån samtidigt som en liten svans av uppgifter är betydligt mer utsatt. Studiens poäng är att integriteten måste kontrolleras på patientnivå, inte bara sammantaget.

Utsattheten är ojämn — och drabbar de underrepresenterade. De patienter som var mest sårbara för en nästan perfekt attack tillhörde systematiskt grupper som var **underrepresenterade i data-**

materialet: etniska minoriteter, fenotyper vid sällsynta sjukdomar och ovanliga bildkarakteristika. I datamängden med elektroniska journaler förekom svarta patienter **31 % oftare** än väntat bland de mest sårbara uppgifterna; i mammografimaterialet var bilder som flaggats som misstänkta för malignitet överrepresenterade i denna riskzon med hela **+1 179 %**. (Dessa två siffror är exempel, inte hela bilden — studien rapporterar samma snedfördelning i flera grupper — och de är *relativa* överrepresentationer inom den lilla svansen med extrem risk: hur mycket oftare dessa patienter förekommer bland de mest utsatta uppgifterna, inte andelen svarta patienter eller patienter med misstänkta mammografibilder som är utsatta.) En modell har färre liknande exempel som dessa patienter kan smälta in bland, så deras uppgifter sticker ut — och det är precis vad en medlemskapsattack upptäcker. Integritetsbristen är inte slumpmässig; den koncentreras till människor som redan befinner sig i marginalen.

Större modeller gör det värre. Antalet patienter som utsattes för nästan perfekta attacker steg kraftigt med modellkapaciteten — i en datamängd för dermatologi ökade andelen från i stort sett noll i den minsta modellen till omkring **en av tio** i den största. I takt med att medicinska AI-modeller blir större och mer kapabla — den riktning hela fältet rör sig i — blir just denna risk, varnar författarna, allvarligare, inte mindre.

Rückerts sammanfattning är rättfram: ”Det här är ingen acceptabel risk. Hälsodata är mycket känsliga.”

Varför ”säker i genomsnitt” är fel test

Det är frestande att läsa en låg genomsnittlig risk som ett friskintyg. Studiens huvudpoäng är att ett genomsnitt här inte bara är oprecist — det mäter fel sak.

En integritetsgaranti är meningsfull endast om den gäller för den person som löper störst risk, inte för personen i mitten. En modell där 999 av 1 000 patienter inte kan identifieras men där en kan identifieras nästan perfekt är inte ”99,9 % privat” i någon mening som har betydelse för den enda personen — och om den personen förutsägbart är patienten med en sällsynt sjukdom eller patienten från en etnisk minoritet är måttet inte bara ofullständigt, utan diskriminerande i det tysta. Det rapporterar säkerhet för majoriteten och kallar det säkerhet för alla.

Det är den förskjutning studien tvingar fram: från *hur privat är modellen i genomsnitt?* till *vilken patient är mest utsatt, och vem är den personen?* Det är olika frågor, och bara den andra är en integritetsfråga.

Vad detta *inte* bevisar — eller påstår

- Det säger **inte** att medicinsk AI bör överges. Författarnas utgångspunkt är riskminskning, inte reträtt: mät och åtgärda risken, sluta inte bygga.
- Det betyder **inte** att varje medicinsk modell läcker eller att någon viss driftsatt modell har angripits. Det visar att *risken finns och är ojämnt fördelad*, och att vanliga aggregerade mått missar den.

- Det betyder **inte** att dina journaluppgifter redan är röjda. Attacken kräver specifika förutsättningar — tillgång till modellen, en tänkbar journaluppgift och angriparens egen infrastruktur — och är ingen enkel vardagsförmåga.
- Det visar **inte** att patienternas *innehåll* läcker. Det som läcker är medlemskapet — faktumet att uppgiften ingår — vilket är farligt av en annan anledning, inte för att journalen lämnas ut.
- Det reducerar **inte** skillnaderna till en enda rubrikvänlig siffra. Det starka, reproducerbara resultatet är *mönstret* — aggregerade mått underskattar den individuella risken, och underrepresenterade patienter bär den största delen av den — i flera datamängder och hundratals modeller.

Hur starka är beläggen?

Detta är ett empiriskt, metodologiskt resultat, och ett robust sådant. Mönstret — att aggregerade integritetsmått systematiskt underskattar risken per patient, att den kvarvarande risken koncentreras till underrepresenterade grupper och att den förvärras med modellkapaciteten — återkom i **sju datamängder med olika datatyper** och ett stort antal modeller per datamängd. Det är just denna bredd som gör mätresultatet trovärdigt, i stället för att det framstår som en artefakt från en enda datamängd.

Två uppriktiga förbehåll. För det första är detta en demonstration av *risk*, uppmätt genom att köra de attacker som författarna själva byggde; den beskriver hur väl en kapabel angripare *skulle kunna* lyckas under angivna antaganden, inte hur ofta verkliga attacker sker. För det andra beskriver de slående siffrorna — nästan perfekt attackframgång för vissa patienter — de värst drabbade individerna *avsiktligt*; det är hela poängen, och de bör läsas som ”svansen är mycket tyngre än genomsnittet antyder”, inte som ”de flesta patienter är utsatta”.

Den rimliga hållningen är varken alarmism eller avfärdande: en omsorgsfull studie av flera datamängder som visar att det vanliga sättet att intyga integritet hos medicinsk AI är blint för sina egna värsta fall — och att dessa värsta fall drabbar de patienter som har minst marginaler.

Varför det spelar roll

Två saker gör detta till mer än en teknisk fotnot.

För det första förändrar det vad ”integritetsbevarande” bör få betyda. Om en modell släpps med ett genomsnittligt integritetsvärde kan detta värde verkligen vara lågt och ändå dölja en delmängd patienter som går att identifiera nästan perfekt med en medlemskapsattack. Studiens praktiska krav är konkret: bedöm integritetsrisken **för varje patient** före lansering, kontrollera vem som kan få tillgång till driftsatta modeller och använd metoder som **differentiell integritet** — små, noggrant kalibrerade störningar som läggs till under träningen och försvagar medlemskapsattacker, till en verklig och hanterad kostnad för modellens användbarhet (avvägningen mellan integritet och användbarhet är uttalad, inte gratis). ”Vi kontrollerade genomsnittet” bör inte längre räknas som att man har kontrollerat.

För det andra flätar studien samman integritet och rättvisa. Samma grupper som är underrepresenterade i medicinska data — och därför redan får sämre hjälp av medicinsk AI — visar sig vara de vars integritet denna AI skyddar minst. Ett fält som arbetar hårt med den första ojämlikheten kan inte behandla den andra som någon annans ansvarsområde. Det är samma människor.

Den betryggande berättelsen om integritet i medicinsk AI — *vi mätte den, genomsnittet är lågt, allt är bra* — är den berättelse studien plockar isär. Inte för att skrämma bort någon från tekniken, utan för att flytta normen dit den hör hemma: till ett löfte som man bara kan hävda att man håller om man har kontrollerat det för den person som löper störst risk att skadas.

Ren sammanfattning

Medlemskapsinferensattacker försöker avgöra om en viss persons uppgifter ingick i en AI-modells träningsdata — och eftersom medlemskapet i sig kan avslöja något (till exempel att personen hade en viss sjukdom) är detta en verklig integritetsläcka även om den underliggande journaluppgiften aldrig kommer fram. Tidigare forskning mätte hur ofta sådana attacker lyckas *i genomsnitt* i en datamängd. Den här studien mätte det *för varje patient* i sju medicinska datamängder (bilddiagnostik, EKG, elektroniska journaler) och många modeller per datamängd, och fann att aggregerade mått kraftigt underskattar risken: vissa individer kan identifieras nästan perfekt (attack-AUC $\geq$ 0,95) även när genomsnittet för hela datamängden ser säkert ut; de mest sårbara kommer systematiskt från underrepresenterade grupper (etniska minoriteter, sällsynt sjukdom, ovanlig bilddiagnostik); och problemet växer med modellstorleken. Författarna argumenterar inte för att överge medicinsk AI — de argumenterar för att mäta integriteten på individnivå, kontrollera tillgången till modeller och använda differentiell integritet. Slutsatsen: ”privat i genomsnitt” är ingen integritetsgaranti, och de som den sviker är vanligtvis de som redan har sämst skydd.

Utan omsvep

Vad studien visar: I sju medicinska datamängder och många modeller per datamängd är risken för medlemskapsinferens per patient mycket högre för vissa individer än aggregerade mått antyder — upp till nästan perfekt attackframgång (AUC $\geq$ 0,95) — och denna kvarvarande risk drabbar oproportionerligt ofta underrepresenterade grupper och växer med modellkapaciteten.

Vad som är rimligt men inte bevisat: Att verkliga angripare redan utnyttjar detta mot driftsatta kliniska modeller; studien visar *förmågan och dess fördelning* under angivna antaganden, inte hur ofta attacker sker i verkligheten.

Vad den inte visar: Att medicinsk AI bör överges; att varje modell läcker eller att någon specifik driftsatt modell har utsatts för intrång; att *innehållet* i patientjournaler (snarare än medlemskapet) röjs; en enda kvantifierad skillnadssiffra — det robusta resultatet är mönstret, inte ett enda tal.

Huvudsakliga begränsningar: Den mäter värstafallsrisk med författarnas egna attacker, inte ob-

serverade incidenter; de dramatiska siffrorna beskriver avsiktligt de mest utsatta individerna; och de exakta skillnaderna mellan grupper visas som ett konsekvent mönster i flera datamängder i stället för att reduceras till en enda siffra.

Hur stort förtroende bör en allmän läsare ha? Stort förtroende för att aggregerade integritetsmått underskattar den individuella risken, att den kvarvarande risken koncentreras till underrepresenterade patienter och att detta förvärras med modellstorleken — det har visats i många datamängder och modeller. Måttligt förtroende när det gäller hur ofta sådana attacker sker i verkligheten, vilket studien inte mäter. Den säkra läsningen är inte ”medicinsk AI läcker dina data” och inte ”integritetsproblemet är löst”, utan ”den vanliga integritetskontrollen är blind för sina egna värsta fall, och dessa fall drabbar de mest sårbara patienterna — därför måste kontrollen förändras.”

Källor

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