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

O AI medicală poate fi privată în medie și totuși să expună anumiți pacienți — mai ales pe cei subreprezentați

Un model de AI medicală numit "privacy-preserving" se sprijină de obicei pe un singur număr mediu. Acest studiu susține că acel număr este testul greșit. Studiind atacuri de membership inference — care arată dacă dosarul unei persoane anume a fost în datele de antrenament ale unui model și astfel poate trăda că persoana a avut o anumită boală — autorii au măsurat riscul per pacient, nu agregat, pe șapte dataseturi medicale (imagistică, ECG, dosare electronice) și multe modele pentru fiecare. Tiparul: modele care par sigure în medie pot totuși lăsa un atacator să identifice aproape perfect indivizi specifici (attack AUC $\geq 0{,}95$); cei expuși sunt sistematic cei din grupuri subreprezentate (etnii minoritare, boli rare, imagistică neobișnuită); iar problema se agravează pe măsură ce modelele cresc. Autorii nu spun să abandonăm AI medicală — spun să măsurăm confidențialitatea per pacient, să controlăm accesul la modele și să folosim differential privacy. Miezul incomod: "privat în medie" nu este o garanție de confidențialitate și îi ratează pe pacienții deja cel mai puțin protejați.

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

O garanție de confidențialitate este o promisiune făcută unei persoane, nu unei medii

Când despre un model de AI medicală se spune că „protejează confidențialitatea", afirmația se sprijină adesea pe un singur număr: pentru ansamblul pacienților ale căror date au fost folosite la antrenare, probabilitatea medie de a deduce dacă o anumită persoană a făcut parte din setul de antrenare este mică. Sună liniștitor. Problema incomodă evidențiată de această lucrare este că media poate fi indicatorul greșit.

Confidențialitatea nu este o medie. Este o promisiune făcută fiecărei persoane: faptul că datele sale au fost incluse în setul de antrenare nu ar trebui să o expună. O medie poate părea liniștitoare pentru aproape toată populația și totuși să ascundă un risc extrem pentru câțiva indivizi. Cercetătorii au măsurat riscul separat pentru fiecare pacient, la o granularitate neobișnuită, și au găsit exact acest tipar: modele care par sigure în ansamblu pot dezvălui aproape perfect apartenența anumitor persoane la setul de antrenare — iar cei mai expuși sunt disproporționat pacienți care aparțin unor grupuri deja slab reprezentate.

Ce au făcut autorii

Echipa — condusă de Moritz Knolle, cu Daniel Rückert (ambii la Technical University of Munich) și Georg Kaissis (Hasso Plattner Institute), alături de colegi de la Imperial College London — a pornit de la o amenințare bine cunoscută, numită **atac de inferare a apartenenței** (*membership inference attack*), și a schimbat întrebarea. În loc să întrebe „în medie, cât de des reușește atacul în setul de date?", cercetătorii au întrebat: „pentru *acest pacient anume*, cât de mare este riscul?"

Au aplicat analiza la nivel individual pe **șapte seturi de date medicale consacrate**, care acoperă tipuri foarte diferite de date — imagistică medicală, electrocardiograme și dosare electronice de sănătate — folosind câte 200 de modele pentru fiecare set. Pentru fiecare pacient și fiecare model, au estimat cu câtă certitudine ar putea un atacator să stabilească dacă dosarul acelei persoane fusese folosit la antrenare, apoi au analizat rezultatele pe grupuri: statutul bolii, rasa auto-raportată, tipul de asigurare, sexul și protocolul imagistic.

> **Ce este, de fapt, un atac de inferare a apartenenței**
>
> Scurgerea de informații este mai subtilă decât situația în care „modelul îți reproduce dosarul". Este vorba despre *apartenență* — simplul fapt că datele tale au fost incluse în setul de antrenare. Începe cu ceea ce unul dintre aceste modele face în mod normal. Îi dai o scanare — o radiografie toracică, să zicem — și îți dă înapoi probabilități: 78% șansă de pneumonie, alături de citiri pentru cardiomegalie, edem, consolidare. Acel răspuns diagnostic de zi cu zi este *singurul* lucru pe care atacul îl folosește. Nimic exotic.
> Slăbiciunea pe care se sprijină este aceasta: un model este de obicei puțin **mai încrezător** pe exemplele exacte pe care a fost antrenat decât pe cele pe care nu le-a văzut niciodată. Gândește-

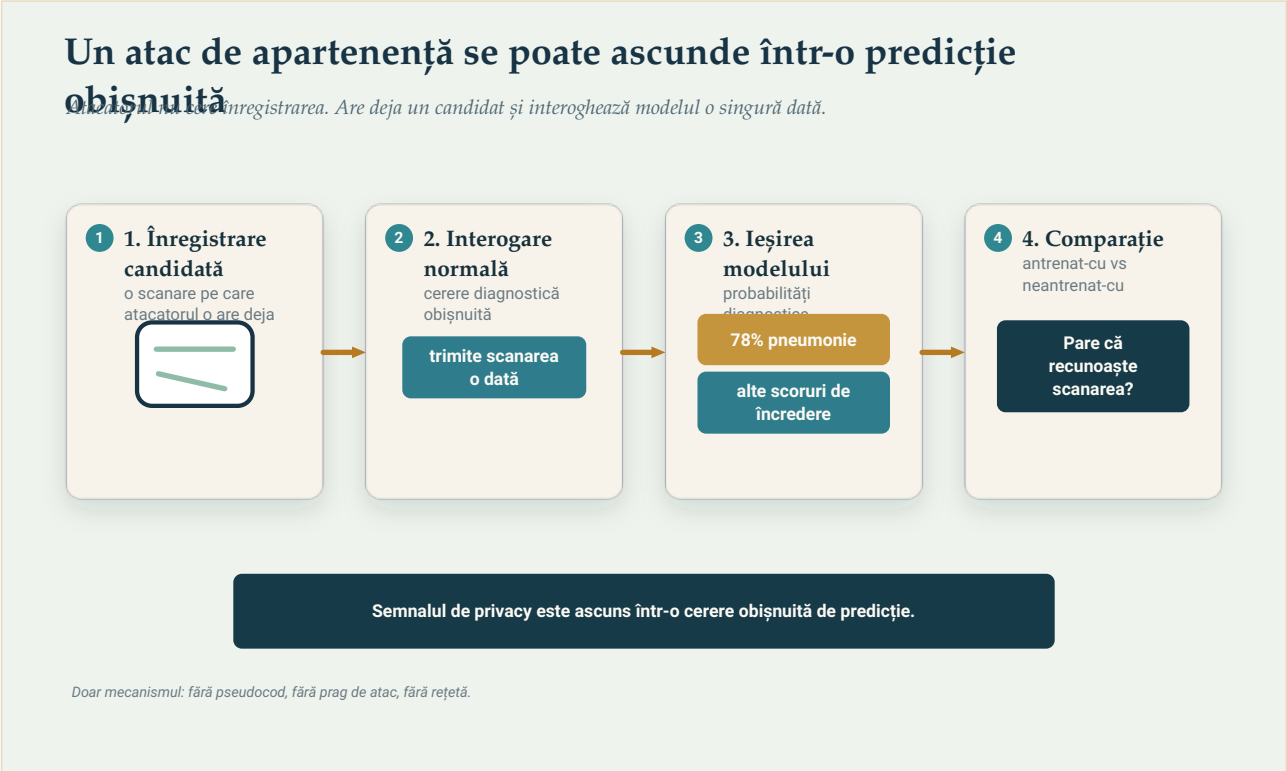
te la un student care a văzut pe ascuns examenul înainte — la întrebările pe care le-a exersat, răspunsurile vin puțin prea repede, puțin prea sigur. A deduce din acel indiciu dacă un anumit dosar a fost printre exemplele folosite la antrenare este tocmai ceea ce înseamnă „inferarea apartenenței".

Așadar, acesta este atacul, pas cu pas. Cineva vrea să știe dacă scanarea unei anumite persoane a fost folosită pentru a antrena modelul. **Nu** cere modelului dosarul — deține deja o scanare candidată (a persoanei, sau o copie apropiată). O trimite o singură dată, ca o cerere diagnostică obișnuită, și notează cât de încrezător este răspunsul. Apoi verifică dacă nivelul de încredere seamănă mai mult cu cel al unui model care *a fost* antrenat pe scanare sau cu cel al unui model care *nu a fost* antrenat pe ea. Comparația poate fi construită relativ ieftin prin antrenarea unui **model de referință**, care arată cum se manifestă acea încredere neobișnuit de mare, fără hardware specializat. Dacă modelul real este neobișnuit de sigur pe această scanare — ca și cum ar recunoaște-o — aceasta este o dovadă puternică că scanarea a fost în datele de antrenament.

Partea neliniștitoare este că interogarea nu arată ca un atac: este aceeași cerere diagnostică pe care ar putea-o face un clinician, iar semnalul legat de confidențialitate este ascuns într-o predicție obișnuită. Tocmai de aceea riscul este relevant, chiar dacă până acum a fost demonstrat în condiții de laborator explicit definite, nu observat în atacuri reale.

De ce contează apartenența, dacă dosarul însuși nu este dezvăluit niciodată? Pentru că *apartenența este ea însăși o informație*. Dacă un model a fost antrenat pe pacienți care au primit o anumită imunoterapie oncologică, atunci confirmarea că dosarul tău este în el dezvăluie că probabil ai avut acel cancer — genul de lucru pe care un asigurător sau un angajator nu ar trebui să îl poată infera vreodată. Conținutul rămâne sigilat; faptul apartenenței este cel care scapă.

Autorii explică limpede ipotezele — acces la predicțiile obișnuite ale modelului, un dosar candidat și propriul model de referință al atacatorului — nu ca pe un ghid practic, ci pentru ca amenințarea să poată fi evaluată în mod realist.

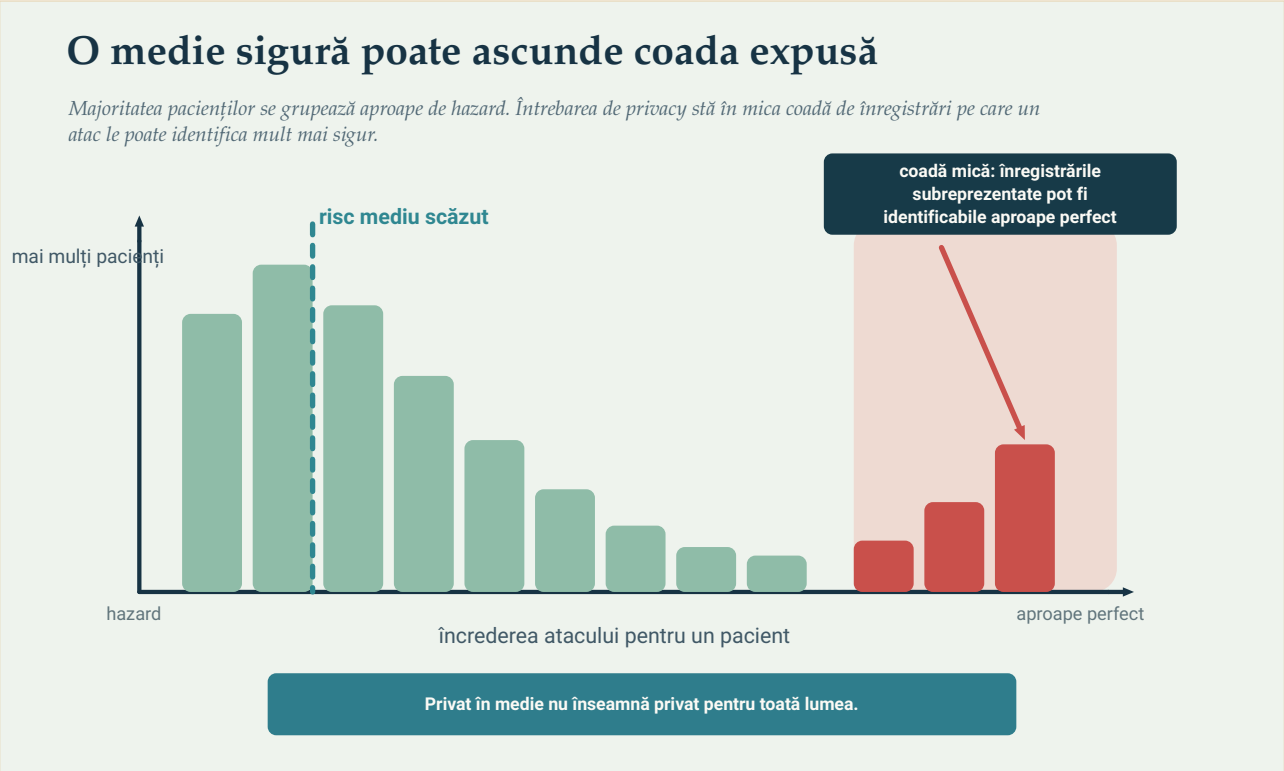


Atacul nu are nevoie de un API special. O interogare diagnostică obișnuită poate dezvălui un semnal de apartenență dacă modelul este neobișnuit de încrezător pentru un dosar pe care l-a văzut în timpul antrenării.

Ce au găsit

Trei constatări, fiecare mai îngrijorătoare decât precedenta.

Mediile îi ascund pe cei mai expuși. Măsurate în ansamblu — abordarea obișnuită — multe dintre aceste modele par să protejeze bine confidențialitatea: pentru majoritatea pacienților, atacul nu se descurcă mai bine decât hazardul. Analiza la nivel de pacient a arătat însă o altă imagine. După cum a spus Knolle în anunțul studiului, evaluările anterioare “au măsurat întotdeauna doar riscul mediu peste toți pacienții. Noi am examinat pentru prima dată riscul la nivelul pacienților individuali — și pictează o imagine foarte diferită.” Pentru unii indivizi, atacul reușește **aproape perfect** — AUC-ul atacului ajunge la **0,95 sau mai mult** (0,5 corespunde hazardului, iar 1,0 unei separări perfecte) — chiar și atunci când media pentru întregul set de date nu pare mai bună decât hazardul.



Media poate sta aproape de șansă în timp ce o mică coadă de dosare rămâne mult mai expusă. Punctul lucrării este că confidențialitatea trebuie verificată la nivelul pacientului, nu doar în agregat.

Expunerea este inegală — și îi afectează mai ales pe cei subreprezentați. Pacienții cei mai vulnerabili la un atac aproape perfect erau sistematic cei din grupuri **subreprezentate în date**: etnii minoritare, fenotipuri de boli rare, caracteristici imagistice neobișnuite. În setul de date de dosare electronice, pacienții din categoria „Black” apăreau **cu 31% mai des** decât era de așteptat printre dosarele cele mai vulnerabile; în setul de mamografii, scanările marcate ca suspecte de malignitate erau suprareprezentate în acea zonă de pericol cu un izbitor **+1.179%**. (Aceste două cifre sunt exemple, nu întreaga imagine — lucrarea raportează aceeași asimetrie în mai multe grupuri — și sunt suprareprezentări *relative* în mica coadă de risc extrem: cât de des apar acești pacienți printre dosarele cele mai expuse, nu fracția pacienților Black sau cu mamografii suspecte care sunt expuși.) Un model are mai puține exemple similare în care să îi estompeze pe acești pacienți, așa că dosarele lor ies în evidență — iar tocmai această particularitate poate fi exploatată de un atac de inferare a apartenenței. Eșecul de confidențialitate nu este aleator; se concentrează pe oamenii deja la margini.

Modelele mai mari înrăutățesc problema. Numărul pacienților expuși la atacuri aproape perfecte a crescut abrupt cu capacitatea modelului — într-un set de date dermatologic, ponderea a urcat de la practic zero în cel mai mic model la aproximativ **unul din zece** în cel mai mare. Pe măsură ce modelele de AI medicală devin mai mari și mai capabile — direcția în care merge întregul domeniu — acest risc specific, avertizează autorii, devine mai sever, nu mai mic.

Rückert rezumă direct concluzia: riscul nu este tolerabil, deoarece datele de sănătate sunt extrem de sensibile.

De ce “sigur în medie” este testul greșit

Tentația este să interpretezi un risc mediu scăzut ca pe un certificat de siguranță. Lecția centrală a lucrării este că media nu este doar imprecisă aici — poate măsura altceva decât riscul care contează pentru individ.

O garanție de confidențialitate are sens doar dacă ține pentru persoana cea mai expusă riscului, nu pentru persoana din mijloc. Un model în care 999 din 1.000 de pacienți sunt nerecuperabili, dar unul poate fi identificat aproape perfect, nu este “99,9% privat” în niciun sens care contează pentru acea persoană — iar dacă acea persoană este previzibil pacientul cu boală rară sau pacientul dintr-o minoritate etnică, metrica nu este doar incompletă, ci discret discriminatorie. Raportează siguranță pentru majoritate și o numește siguranță pentru toți.

Aceasta este schimbarea pe care lucrarea o forțează: de la *cât de privat este acest model în medie?* la *cine este pacientul cel mai expus și cine este el?* Sunt întrebări diferite, și doar a doua este o întrebare de confidențialitate.

Ce *nu* demonstrează — sau susține

- **Nu** spune că AI medicală ar trebui abandonată. Încadrarea autorilor este mitigare, nu retragere: măsoară și repară riscul, nu opri construcția.
- **Nu** înseamnă că fiecare model medical scurge date sau că un anumit model implementat a fost atacat. Arată că *riscul există și este distribuit inegal,* și că metricile agregate standard îl ratează.
- **Nu** înseamnă că dosarele pacienților sunt deja expuse. Atacul necesită condiții specifice — acces la model, un dosar candidat și infrastructura necesară atacatorului — nu este o capacitate care apare automat.
- **Nu** arată că se scurge *conținutul* dosarelor pacienților. Informația care poate fi dedusă este apartenența la setul de antrenare — faptul includerii — care este periculoasă dintr-un alt motiv, nu pentru că dosarul ar fi reprodus.
- **Nu** reduce disparitățile la un singur număr de titlu. Rezultatul puternic și reproductibil este *tiparul* — metricile agregate subestimează riscul individual, iar pacienții subreprezentați poartă cea mai mare parte — peste seturi de date și sute de modele.

Cât de solide sunt dovezile?

Acesta este un rezultat empiric, metodologic, și unul robust. Tiparul — metricile agregate de confidențialitate subestimează sistematic riscul la nivel de pacient, riscul rezidual se concentrează pe grupurile subreprezentate și se agravează cu capacitatea modelului — a ținut pe **șapte seturi de date de tipuri diferite** și pe un număr mare de modele per set de date. Această lărgime este exact ce face credibilă o afirmație de măsurare, nu un artefact al unui singur set de date.

Două rezerve sunt importante. În primul rând, aceasta este o demonstrație de *risc*, măsurată cu atacurile construite de autori; ea arată cât de bine s-ar putea descurca un atacator capabil în ipotezele declarate, nu cât de frecvente sunt atacurile reale. În al doilea rând, cifrele cele mai spectaculoase — succes aproape perfect pentru unii pacienți — descriu în mod intenționat indivizii cei mai expuși. Ele trebuie citite ca „extrema distribuției este mult mai riscantă decât sugerează media", nu ca „majoritatea pacienților sunt expuși".

Poziția potrivită nu este nici alarmă, nici respingere: un studiu atent, multi-set de date, care arată că modul standard în care certificăm confidențialitatea AI medicale este orb la propriile cazuri cele mai rele — iar acele cazuri cele mai rele cad pe pacienții cu cea mai mică marjă de pierdut.

De ce contează

Două lucruri fac asta mai mult decât o notă tehnică.

În primul rând, schimbă ce ar trebui să însemne „protejează confidențialitatea". Un model poate avea un scor mediu foarte bun și totuși să ascundă un subset de pacienți aproape perfect identificabili printr-un atac de inferare a apartenenței. Recomandarea practică a lucrării este concretă: evaluați riscul **pentru fiecare pacient** înainte de lansare, controlați accesul la modelele implementate și folosiți tehnici precum **confidențialitatea diferențială** (*differential privacy*) — introducerea atent calibrată a unui nivel de zgomot în timpul antrenării, care reduce eficiența acestor atacuri cu un cost real, dar controlabil, asupra utilității modelului. Compromisul dintre confidențialitate și utilitate este explicit; „am verificat media" nu ar trebui să mai fie suficient.

În al doilea rând, leagă problema confidențialității de cea a echității. Aceleași grupuri care sunt subreprezentate în datele medicale — și deci deja servite mai prost de AI medicală — se dovedesc a fi cele a căror confidențialitate această AI o protejează cel mai puțin. Un domeniu care muncește mult la prima inechitate nu poate trata a doua ca pe departamentul altcuiva. Sunt aceiași oameni.

Povestea liniștitoare a confidențialității în AI medicală — *am măsurat-o, media este scăzută, suntem bine* — este povestea pe care această lucrare o demontează. Nu ca să sperie pe cineva de tehnologie, ci ca să mute standardul acolo unde îi este locul: o promisiune pe care poți spune că o păstrezi doar dacă ai verificat-o pentru persoana cea mai probabilă să fie rănită.

Pe scurt

Atacurile de inferare a apartenenței încearcă să determine dacă dosarul unei anumite persoane a fost inclus în datele de antrenare ale unui model AI — iar faptul apartenenței poate fi el însuși revelator (că ai avut o anumită boală, de exemplu), aceasta este o scurgere reală de confidențialitate chiar când dosarul de bază nu apare niciodată. Lucrările anterioare măsurau cât de des reușesc astfel de atacuri *în medie* peste un set de date. Acest studiu a măsurat riscul *la nivel de pacient*, pe șapte seturi de date medicale (imagistică, ECG, dosare electronice) și multe modele fiecare, și a găsit că metricile

agregate subestimează grav riscul: unii indivizi pot fi identificați aproape perfect (AUC al atacului $\geq$ 0,95) chiar când media întregului set de date pare sigură; cei mai vulnerabili sunt sistematic cei din grupuri subreprezentate (etnii minoritare, boli rare, imagistică neobișnuită); iar problema crește cu dimensiunea modelului. Autorii nu cer abandonarea AI medicale — cer măsurarea confidențialității la nivel individual, controlul accesului la modele și folosirea confidențialității diferențiale. Concluzia: "privat în medie" nu este o garanție de confidențialitate, iar oamenii pe care îi ratează sunt de obicei cei deja cel mai puțin protejați.

Verificare fără exagerări

Ce arată lucrarea: Pe șapte seturi de date medicale și multe modele fiecare, riscul individual de inferare a apartenenței este mult mai mare pentru unii indivizi decât sugerează metricile agregate — până la succes aproape perfect al atacului (AUC $\geq$ 0,95) — iar acel risc rezidual cade disproporționat pe grupuri subreprezentate și crește cu capacitatea modelului.

Ce este plauzibil, dar nedovedit: Că atacatori reali exploatează deja asta împotriva modelelor clinice implementate; studiul demonstrează *capacitatea și distribuția ei* sub ipoteze declarate, nu frecvența atacurilor în utilizare reală.

Ce nu arată: Că AI medicală ar trebui abandonată; că fiecare model scurge date sau că un anumit model implementat a fost compromis; că *conținutul* dosarelor pacienților (mai degrabă decât apartenența la setul de antrenare) este expus; o singură cifră de disparitate — rezultatul robust este tiparul, nu un număr.

Limitări principale: Măsoară riscul din scenariul cel mai nefavorabil prin atacurile autorilor, nu incidente observate; cifrele dramatice descriu prin design indivizii cei mai expuși; iar disparitățile exacte pe grup sunt arătate ca un tipar consistent peste seturi de date, nu reduse la un singur număr.

Câtă încredere ar trebui să aibă un cititor general? Mare că metricile agregate de confidențialitate subestimează riscul individual, că riscul rezidual se concentrează pe pacienții subreprezentați și că se agravează cu dimensiunea modelului — este demonstrat pe multe seturi de date și modele. Moderată privind frecvența reală a acestor atacuri, pe care acest studiu nu o măsoară. Lectura sigură: nu "AI medicală îți scurge datele" și nu "confidențialitatea este rezolvată", ci "verificarea standard de confidențialitate este oarbă la propriile cazuri cele mai rele, iar acele cazuri cad pe cei mai vulnerabili pacienți — deci verificarea trebuie să se schimbe."

Surse

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