The Clean Paper

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

Një AI mjekësore mund të jetë private mesatarisht dhe megjithatë të ekspozojë pacientë të caktuar — sidomos të nënpërfaqësuarit

Pretendimi se një model i AI-së mjekësore "ruan privatësinë" zakonisht mbështetet në një numër mesatar. Ky studim argumenton se ai është testi i gabuar. Duke studiuar sulme për inferimin e anëtarësisë — që zbulojnë nëse kartela e një personi ishte në të dhënat e trajnimit dhe, kështu, mund të tregojnë se ai kishte një sëmundje të caktuar — autorët matën rrezikun për çdo pacient, jo vetëm në agregat, në shtatë dataset-e mjekësore dhe shumë modele. Modelet që duken të sigurta mesatarisht mund të lejojnë identifikim pothuajse të përsosur të individëve të caktuar (AUC $\geq$ 0.95); të ekspozuarit janë sistematikisht nga grupet e nënpërfaqësuara; dhe rreziku rritet me madhësinë e modelit. Autorët nuk kërkojnë braktisjen e AI-së mjekësore, por matje individuale të privatësisë, kontroll të aksesit dhe privatësi diferenciale.

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

Garancia e privatësisë është një premtim ndaj një personi, jo ndaj mesatares

Kur një model i AI-së mjekësore quhet "që ruan privatësinë", ky pretendim zakonisht mbështetet te një numër: në tërësinë e pacientëve, të dhënat e të cilëve u përdorën për trajnimin e modelit, probabiliteti mesatar që të mund të zbulohet nëse një individ ishte pjesë e grupit të trajnimit është i ulët. Tingëllon qetësuese. Pika e qetë, por e pakëndshme, e këtij punimi është se ky është numri i gabuar.

Privatësia nuk është mesatare. Është një premtim ndaj çdo individi — se fakti që të dhënat e tij ishin në grupin e trajnimit nuk do të kthehet më vonë në një mënyrë për ta ekspozuar. Dhe një mesatare mund ta mbajë këtë premtim për pothuajse të gjithë, ndërsa ta thyejë plotësisht për disa. Studiuesit matën rrezikun e privatësisë pacient pas pacienti, me një imtësi që nuk ishte përdorur më parë, dhe gjetën pikërisht këtë: modele që duken të sigurta në nivel të përgjithshëm mund të zbulojnë pothuajse në mënyrë të përsosur anëtarësinë e individëve të caktuar — dhe këta individë janë, në mënyrë disproporcionale, ata që tashmë janë më pak të mbrojtur.

Çfarë bënë autorët

Ekipi — i udhëhequr nga Moritz Knolle, me Daniel Rückert (të dy në Technical University of Munich) dhe Georg Kaissis (Hasso Plattner Institute), së bashku me kolegë nga Imperial College London — mori një kërcënim të njohur të privatësisë, të quajtur **sulm për inferimin e anëtarësisë** (*membership inference attack*), dhe ndryshoi pyetjen. Në vend që të pyesnin "mesatarisht, sa shpesh ka sukses ky sulm në të gjithë dataset-in?", pyetën: "për *këtë pacient të caktuar*, sa i ekspozuar është ai?"

Ata e kryen këtë analizë për çdo pacient në **shtatë dataset-e të konsoliduara mjekësore**, me lloje shumë të ndryshme të dhënash — imazheri mjekësore, elektrokardiograme dhe kartela elektronike shëndetësore — dhe, në secilin prej tyre, në 200 modele. Për çdo pacient në çdo model, vlerësuan se me çfarë sigurie një sulmues mund të dallonte nëse kartela e atij personi kishte qenë pjesë e të dhënave të trajnimit, pastaj i ndanë rezultatet sipas grupeve: status sëmundjeje, racë të vetëdeklaruar, sigurim, seks dhe protokoll imazherie.

> **Çfarë është në të vërtetë një sulm për inferimin e anëtarësisë**
>
> Rrjedhja këtu është më e hollë se "modeli nxjerr kartelën tënde". Bëhet fjalë për *anëtarësinë* — vetëm faktin që të dhënat e tua ishin në grupin e trajnimit.
> Nisemi nga ajo që bën normalisht një model i tillë. I jep një skanim — për shembull një radiografi toraksi — dhe ai kthen probabilitete: 78% mundësi për pneumoni, së bashku me vlerësime për kardiomegali, edemë, konsolidim. Pikërisht kjo përgjigje e zakonshme diagnostike është *e vetmja* gjë që përdor sulmi. Asgjë ekzotike.
> Dobësia mbi të cilën mbështetet është kjo: një model zakonisht është pak **më i sigurt** për shembujt e saktë me të cilët është trajnuar sesa për shembuj që nuk i ka parë kurrë. Mendoni

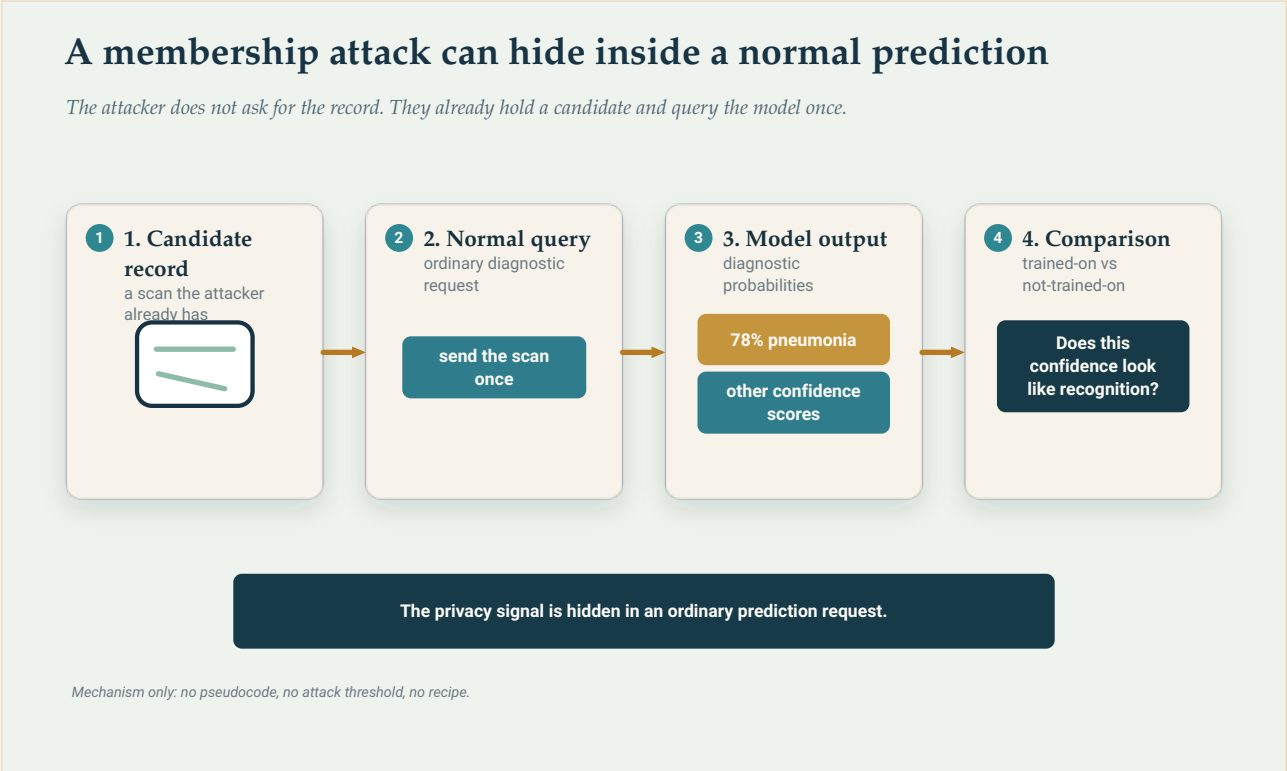
për një student që e ka parë fshehurazi provimin më parë — te pyetjet që i ka ushtruar, përgjigjet vijnë pak tepër shpejt dhe pak tepër me siguri. Të kuptosh nga kjo shenjë nëse një kartelë e caktuar ishte një nga shembujt që modeli studioi, pikërisht kjo do të thotë "inferim i anëtarësisë".

Pra sulmi funksionon kështu. Dikush dëshiron të dijë nëse skanimi i një personi të caktuar u përdor për trajnimin e modelit. Ai **nuk** i kërkon modelit kartelën — tashmë ka një skanim kandidat (të personit ose një kopje shumë të afërt). E dërgon një herë, si kërkesë të zakonshme diagnostike, dhe shënon sa e sigurt është përgjigjja. Pastaj kontrollon nëse kjo siguri i ngjan më shumë sjelljes së një modeli që *ishte* trajnuar me atë skanim apo të një modeli që *nuk ishte* — krahasim që mund ta bëjë me kosto të ulët duke trajnuar një model zëvendësues të vetin ("model referencë") për të mësuar se si duket ajo mbisiguri karakteristike, në një kompjuter të zakonshëm pa pajisje të posaçme. Nëse modeli real është jashtëzakonisht i sigurt për këtë skanim — sikur ta njohë — kjo është provë e fortë se skanimi ishte në të dhënat e trajnimit.

Pjesa shqetësuese është se kërkesa nuk duket aspak si sulm: është e njëjta pyetje diagnostike që do të bënte një klinicist, dhe sinjali i privatësisë fshihet brenda një parashikimi të zakonshëm. Pikërisht kjo e bën rrezikun konkret — edhe pse deri tani është demonstruar nën supozime të deklaruara laboratorike, jo kapur në përdorim real.

Pse ka rëndësi anëtarësia, nëse vetë kartela nuk del kurrë? Sepse *anëtarësia është fakt*. Nëse një model është trajnuar mbi pacientë që morën një imunoterapi të caktuar kundër kancerit, konfirmimi se kartela jote është në të zbulon se me gjasë ke pasur atë kancer — diçka që një sigurues apo punëdhënës nuk duhet të jetë në gjendje ta nxjerrë si përfundim. Përmbajtja mbetet e mbyllur; ajo që del jashtë është fakti i përkatësisë.

Autorët i paraqesin hapur këto supozime — akses te parashikimet e zakonshme të modelit, një kartelë kandidate dhe modelin e referencës të vetë sulmuesit — jo si udhëzim praktik, por që kërcënimi të mund të analizohet konkretisht, jo të trajtohet me hamendësime.



Sulmi nuk ka nevojë për ndonjë API të posaçme privatësie. Një kërkesë normale diagnostike mund të nxjerrë një sinjal anëtarësie nëse modeli është jashtëzakonisht i sigurt për një kartelë që e ka parë më parë.

Original Aurora diagram — The Clean Paper CC BY 4.0

Çfarë gjetën

Tre rezultate, secili më i mprehtë se tjetri.

Mesataret fshehin të ekspozuarit. Kur maten në agregat — mënyra e zakonshme — shumë nga këto modele duken qetësueshëm private: për shumicën e pacientëve, sulmi nuk bën më mirë se rastësia. Pamja pacient pas pacienti tregoi një histori tjetër. Siç tha Knolle në njoftimin e studimit, vlerësimet e mëparshme “kanë matur gjithmonë vetëm rrezikun mesatar për të gjithë pacientët. Ne e shqyrtuam për herë të parë rrezikun në nivel të pacientëve individualë — dhe kjo jep një pamje shumë të ndryshme.” Për disa individë, sulmi ka sukses **pothuajse të përsosur** — autorët e matin me AUC të sulmit **0.95 ose më të lartë** (0.5 është si hedhja e monedhës, 1.0 është e përsosur) — edhe kur mesatarja e gjithë dataset-it dukej jo më e mirë se rastësia.

Mesatarja mund të jetë pranë nivelit të rastësisë ndërsa një bisht i vogël kartelash mbetet shumë më i ekspozuar. Pika e punimit është se privatësia duhet kontrolluar në nivel pacienti, jo vetëm në agregat.

Original Aurora diagram — The Clean Paper CC BY 4.0

Ekspozimi është i pabarabartë — dhe bie mbi grupet e nënpërfaqësuara. Pacientët më të cenueshëm ndaj një sulmi pothuajse të përsosur ishin në mënyrë sistematike ata nga grupe **të nënpërfaqësuara në të dhëna**: pakica etnike, fenotipe të sëmundjeve të rralla, karakteristika të pazakonta imazherie. Në dataset-in e kartelave elektronike, pacientët Black shfaqeshin **31% më shpesh** se sa pritej mes kartelave më të cenueshme; në dataset-in e mamografive, skanimet e shënuara si të dyshimta për malignitet ishin të mbipërfaqësuara në atë zonë rreziku me një shifër të jashtëzakonshme **+1,179%**. (Këto dy shifra janë shembuj, jo e gjithë tabloja — punimi raporton të njëjtën pririe në disa grupe — dhe janë *mbipërfaqësime relative* brenda bishtit të vogël me rrezik ekstrem: tregojnë sa më shpesh shfaqen këta pacientë mes kartelave më të ekspozuara, jo përqindjen e të gjithë pacientëve Black apo të të gjitha mamografive të dyshimta që janë të ekspozuara.) Modeli ka më pak shembuj të ngjashëm me të cilët t'i "shkrijë" këta pacientë, ndaj kartelat e tyre bien më shumë në sy — dhe pikërisht kjo është ajo që zbulon një sulm anëtarësie. Dështimi i privatësisë nuk është rastësor; përqendrohet te njerëzit që tashmë janë në margjina.

Modelet më të mëdha e përkeqësojnë problemin. Numri i pacientëve të ekspozuar ndaj sulmeve pothuajse të përsosura u rrit ndjeshëm me kapacitetin e modelit — në një dataset dermatologjik, përqindja u ngrit nga pothuajse zero në modelin më të vogël në rreth **një në dhjetë** në më të madhin. Ndërsa modelet e AI-së mjekësore bëhen më të mëdha dhe më të afta — drejtimi ku po lëviz e gjithë fusha — ky rrezik i veçantë, paralajmërojnë autorët, bëhet më serioz, jo më i vogël.

Përmbledhja e Rückert është e prerë: "Ky nuk është rrezik i tolerueshëm. Të dhënat shëndetësore janë shumë të ndjeshme."

Pse “i sigurt mesatarisht” është testi i gabuar

Është joshëse që një rrezik mesatar i ulët të lexohet si certifikatë pastërtie. Mësimi qendror i punimit është se këtu mesatarja nuk është thjesht jo e saktë — ajo mat gjënë e gabuar.

Një garanci privatësie ka kuptim vetëm nëse vlen për personin me rrezikun më të madh, jo për personin në mes. Një model ku 999 nga 1,000 pacientë nuk mund të identifikohen, por njëri mund të identifikohet pothuajse në mënyrë të përsosur, nuk është “99.9% privat” në asnjë kuptim që ka rëndësi për atë një person — dhe nëse ai person është në mënyrë të parashikueshme pacient me sëmundje të rrallë ose pjesëtar i një pakice etnike, metri nuk është vetëm i paplotë, por edhe në heshtje diskriminues. Raporton siguri për shumicën dhe e quan atë siguri për të gjithë.

Ky është ndryshimi që kërkon punimi: nga *sa privat është ky model mesatarisht?* te *kush është pacienti më i ekspozuar dhe kush është ai?* Janë pyetje të ndryshme, dhe vetëm e dyta është vërtet pyetje privatësie.

Çfarë *nuk* provon kjo — ose pretendon

- **Nuk** thotë se AI mjekësore duhet braktisur. Korniza e autorëve është zbutje e rrezikut, jo tërheqje: mateni dhe korrigjojeni rrezikun, mos ndaloni së ndërtuari.
- **Nuk** do të thotë se çdo model mjekësor po rrjedh të dhëna, ose se ndonjë model i caktuar në përdorim është sulmuar. Tregon se *rreziku ekziston dhe shpërndahet në mënyrë të pabarabartë,* dhe se metrikat standarde agregate nuk e shohin.
- **Nuk** do të thotë se kartelat tuaja tashmë janë ekspozuar. Sulmi kërkon kushte specifike — akses te modeli, një kartelë kandidate dhe infrastrukturën e vet të sulmuesit — jo një aftësi rastësore.
- **Nuk** tregon se rrjedh *përmbajtja* e kartelës së pacientit. Ajo që rrjedh është anëtarësia — fakti i përfshirjes — e cila është e rrezikshme për një arsye tjetër, jo sepse kartela del e plotë.
- **Nuk** i redukton pabarazitë në një numër të vetëm për titull. Rezultati i fortë dhe i riprodhueshëm është *modeli i përgjithshëm*: metrikat agregate nënvlerësojnë rrezikun individual dhe pacientët e nënpërfaqësuar mbajnë pjesën më të madhe të tij, në disa dataset-e dhe qindra modele.

Sa e fortë është prova?

Ky është rezultat empirik e metodologjik, dhe është i fortë. Modeli — metrikat agregate të privatësisë nënvlerësojnë në mënyrë sistematike rrezikun pacient për pacient, rreziku i mbetur përqendrohet te grupet e nënpërfaqësuara dhe përkeqësohet me kapacitetin e modelit — u përsërit në **shtatë dataset-e me lloje të ndryshme të dhënash** dhe në një numër të madh modelesh për secilin. Pikërisht kjo gjerësi e bën pretendimin për matjen bindës, jo një artefakt të një dataset-i të vetëm.

Dy kufizime të ndershme. Së pari, kjo është demonstrim i *rrezikut*, i matur duke ekzekutuar sulmet që vetë autorët ndërtuan; karakterizon se sa mirë *mund* të veprojë një sulmues i aftë nën supozimet e dhëna, jo sa shpesh ndodhin sulme reale. Së dyti, shifrat më dramatike — sukses pothuajse i përsosur

për disa pacientë — përshkruajnë individët më të ekspozuar *me qëllim të dizajnit të analizës*; pikërisht kjo është çështja, dhe duhen lexuar si "bishti është shumë më i rëndë se sa lë të kuptohet mesatarja", jo "shumica e pacientëve janë të ekspozuar".

Qëndrimi i duhur nuk është as alarmizëm, as shpërfillje: është një studim i kujdesshëm në shumë dataset-e që tregon se mënyra standarde me të cilën certifikojmë privatësinë e AI-së mjekësore është e verbër ndaj rasteve të veta më të këqija — dhe se këto raste bien mbi pacientët me më pak hapësirë për të përballuar një dështim tjetër.

Pse ka rëndësi

Dy gjëra e bëjnë këtë më shumë se një shënim teknik.

Së pari, ndryshon çfarë duhet të lejohet të nënkuptojë "ruan privatësinë". Nëse një model publikohet me një rezultat mesatar privatësie, ai rezultat mund të jetë vërtet i ulët dhe megjithatë të fshehë një nën-grup pacientësh që mund të identifikohen pothuajse në mënyrë të përsosur nga një sulm anëtarësie. Kërkesa praktike e punimit është konkrete: vlerësoni rrezikun e privatësisë **për çdo pacient** para publikimit, kontrolloni se kush mund të ketë akses te modelet e vendosura dhe përdorni teknika si **privatësia diferenciale** — zhurmë e vogël, e kalibruar me kujdes, e shtuar gjatë trajnimit për të dobësuar sulmet e anëtarësisë, me një kosto reale dhe të menaxhueshme për dobishmërinë e modelit (kompromisi privatësi–dobishmëri është i qartë, jo falas). "Kontrolluam mesataren" nuk duhet të llogaritet më si kontroll i privatësisë.

Së dyti, punimi e lidh privatësinë me drejtësinë. Të njëjtat grupe që janë të nënpërfaqësuara në të dhënat mjekësore — dhe prandaj tashmë shërbehen më keq nga AI mjekësore — rezultojnë të jenë edhe ato privatësinë e të cilave AI e mbron më pak. Një fushë që punon fort mbi pabarazinë e parë nuk mund ta trajtojë të dytën si problem të dikujt tjetër. Janë të njëjtët njerëz.

Historia qetësuese e privatësisë së AI-së mjekësore — *e matëm, mesatarja është e ulët, pra jemi në rregull* — është pikërisht historia që ky punim çmonton. Jo për të trembur njerëzit nga teknologjia, por për ta çuar standardin atje ku duhet: te një premtim që mund të pretendosh se e mban vetëm nëse e ke kontrolluar për personin që ka më shumë gjasa të dëmtohet.

Përmbledhje e shkurtër

Sulmet për inferimin e anëtarësisë përpiqen të përcaktojnë nëse kartela e një personi të caktuar ishte në të dhënat e trajnimit të një modeli AI — dhe, sepse vetë anëtarësia mund të zbulojë informacion (për shembull se personi kishte një sëmundje të caktuar), kjo është rrjedhje reale privatësie edhe kur kartela bazë nuk shfaqet kurrë. Puna e mëparshme mati sa shpesh këto sulme kanë sukses *mesatarisht* në një dataset. Ky studim e mati rrezikun *për çdo pacient*, në shtatë dataset-e mjekësore (imazheri, ECG, kartela elektronike) dhe shumë modele për secilin, dhe zbuloi se metrikat agregate e nënvlerësojnë rëndë rrezikun: disa individë mund të identifikohen pothuajse në mënyrë të përsosur

(AUC i sulmit ≥ 0.95) edhe kur mesatarja e dataset-it duket e sigurt; më të cenueshmit janë sistematikisht ata nga grupe të nënpërfaqësuara (pakica etnike, sëmundje të rralla, imazheri e pazakontë); dhe problemi rritet me madhësinë e modelit. Autorët nuk argumentojnë për braktisjen e AI-së mjekësore — kërkojnë matje të privatësisë në nivel individual, kontroll të aksesit te modeli dhe përdorim të privatësisë diferenciale. Përfundimi: "privat mesatarisht" nuk është garanci privatësie, dhe ata që dështon të mbrojë janë zakonisht pacientët tashmë më pak të mbrojtur.

Kontroll pa teprime

Çfarë tregon punimi: Në shtatë dataset-e mjekësore dhe shumë modele për secilin, rreziku individual nga membership inference është shumë më i lartë për disa persona sesa sugjerojnë metrikat agregate — deri në sukses pothuajse të përsosur të sulmit (AUC ≥ 0.95) — dhe ky rrezik i mbetur bie në mënyrë disproporcionale mbi grupet e nënpërfaqësuara dhe rritet me kapacitetin e modelit.

Çfarë është e besueshme, por jo e provuar: Se sulmues realë tashmë po e shfrytëzojnë këtë kundër modeleve klinike të vendosura; studimi demonstron *aftësinë dhe shpërndarjen e saj* nën supozime të deklaruara, jo shpeshtësinë e sulmeve reale.

Çfarë nuk tregon: Se AI mjekësore duhet braktisur; se çdo model rrjedh të dhëna ose se ndonjë model i caktuar është kompromet uar; se ekspozohet *përmbajtja* e kartelës së pacientit (në vend të anëtarësisë); ose një shifër të vetme që përmbledh pabarazinë — rezultati robust është modeli i përsëritur, jo një numër i vetëm.

Kufizimet kryesore: Mat rrezik të rastit më të keq përmes sulmeve të autorëve, jo incidente të vëzhguara; shifrat dramatike përshkruajnë qëllimisht individët më të ekspozuar; dhe dallimet e sakta sipas grupeve paraqiten si një model i qëndrueshëm në dataset-e, jo si një numër i vetëm.

Sa besim duhet të ketë një lexues i përgjithshëm? Të lartë se metrikat agregate të privatësisë e nënvlerësojnë rrezikun individual, se rreziku i mbetur përqendrohet te pacientët e nënpërfaqësuar dhe se përkeqësohet me madhësinë e modelit — kjo demonstrohet në shumë dataset-e dhe modele. Më të moderuar për shpeshtësinë reale të këtyre sulmeve, të cilën ky studim nuk e mat. Leximi i sigurt nuk është "AI mjekësore nxjerr të dhënat e tua", as "privatësia është zgjidhur", por: "kontrolli standard i privatësisë nuk i sheh rastet e veta më të këqija, dhe këto raste bien mbi pacientët më të cenueshëm — prandaj kontrolli duhet të ndryshojë."

Burimet

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