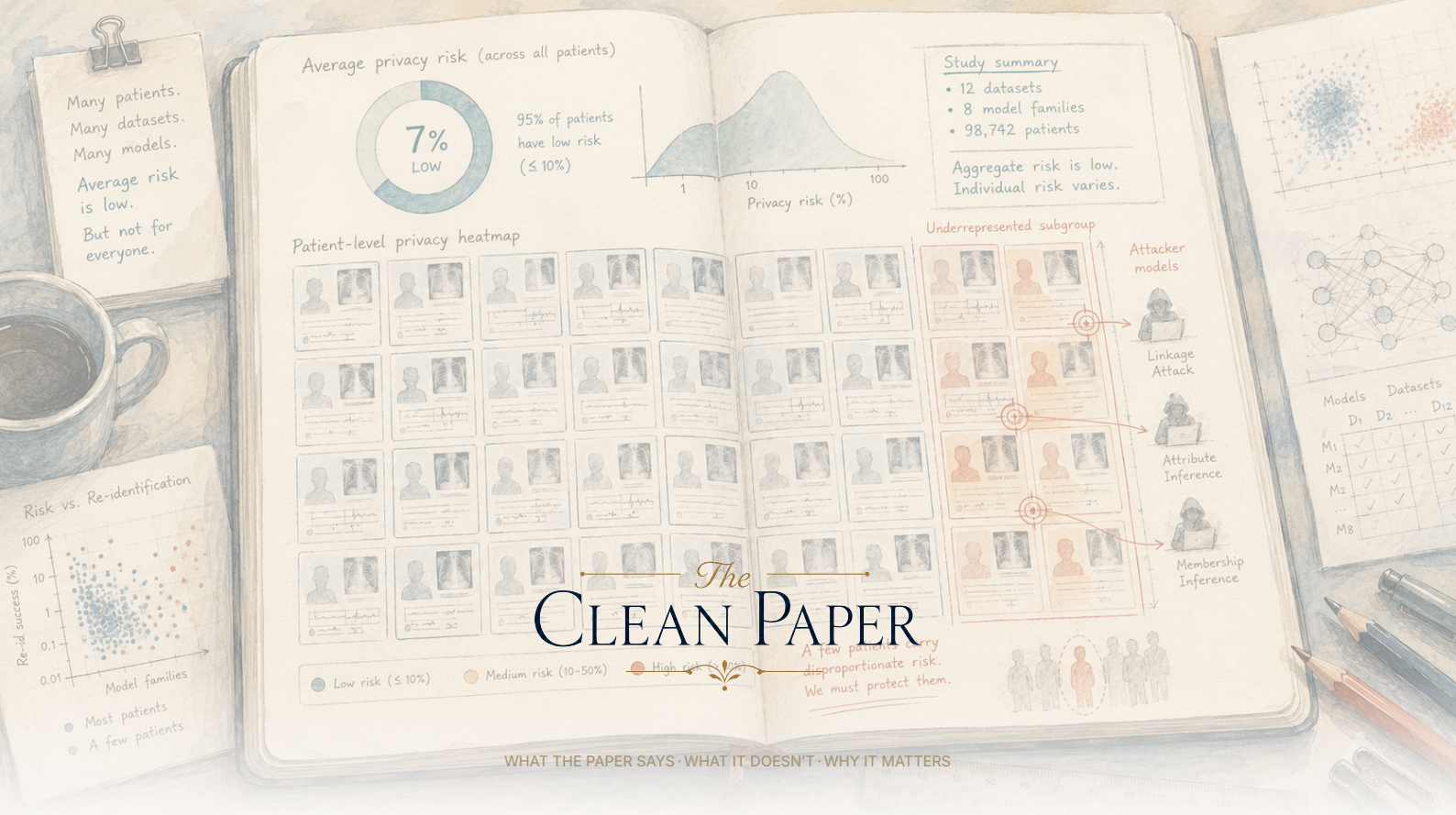


Медицински AI може да биде приватен во просек, а сепак да изложува одредени пациенти — најмногу недоволно застапените

Ознаката „зачувува приватност“ за медицински AI обично се потпира на една просечна бројка. Оваа студија тврди дека тоа е погрешниот тест. Анализирајќи membership inference напади — кои откриваат дали записот на конкретно лице бил во податоците за обука и така може индиректно да открие дека лицето имало одредена болест — авторите го измериле ризикот по пациент, наместо агрегатно, низ седум медицински datasets (снимки, ECG и електронски записи) и многу модели во секој. Образецот е јасен: модели што изгледаат безбедни во просек сепак може да дозволат речиси совршена идентификација на конкретни поединци (attack AUC ≥ 0,95); изложените систематски потекнуваат од недоволно застапени групи — етнички малцинства, ретки болести, невообичаени снимки — а проблемот се влошува со растот на моделите. Авторите не велат да се напушти медицинскиот AI; бараат приватноста да се мери по пациент, пристапот до моделите да се контролира и да се користи диференцијална приватност. Непријатната суштина: „приватно во просек“ не е гаранција за приватност, а најчесто ги изневерува пациентите што и инаку се најмалку заштитени.

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

Гаранцијата за приватност е ветување кон човек, не кон просек

Кога за некој медицински AI-модел се вели дека „ја зачувува приватноста“, тврдењето обично се потпира на една бројка: ако се земат сите пациенти чии податоци биле користени за обука, просечната веројатност да се утврди дали конкретно лице било во тренинг-сетот е мала. Тоа звучи смирувачки. Тивката и непријатна поента на овој труд е дека тоа е погрешната бројка.

Приватноста не е просек. Таа е ветување кон секој човек — дека фактот што неговите податоци биле во тренинг-сетот нема подоцна да го изложи. Просекот може да го одржи тоа ветување за речиси сите, а целосно да го прекрши за неколкумина. Истражувачите го измериле ризикот за приватност пациент по пациент, со резолуција што претходно не била користена, и го нашле токму тоа: модели што изгледаат безбедни во агрегат може речиси совршено да откријат дали одредени луѓе биле во податоците за обука — а тие луѓе непропорционално често се токму оние што и инаку се најслабо заштитени.

Што направиле авторите

Тимот — предводен од Moritz Knolle, со Daniel Rückert (двајцата од Technical University of Munich) и Georg Kaissis (Hasso Plattner Institute), заедно со колеги од Imperial College London — земал добро позната закана за приватноста наречена **membership inference attack** и го сменил прашањето. Наместо „во просек, колку често нападот успева низ целиот dataset?“, тие прашале: „за *овој конкретен пациент*, колку е изложен?“

Таа анализа по пациент ја спровеле врз **седум утврдени медицински datasets**, со многу различни типови податоци — медицински снимки, електрокардиограми и електронски здравствени досиеја — и со по 200 модели во секој dataset. За секој пациент во секој модел процениле колку сигурно напаѓач би можел да утврди дали записот на тоа лице бил дел од податоците за обука, а потоа резултатите ги разделиле според групи: статус на болест, самопријавена раса, осигурување, пол и протокол за снимање.

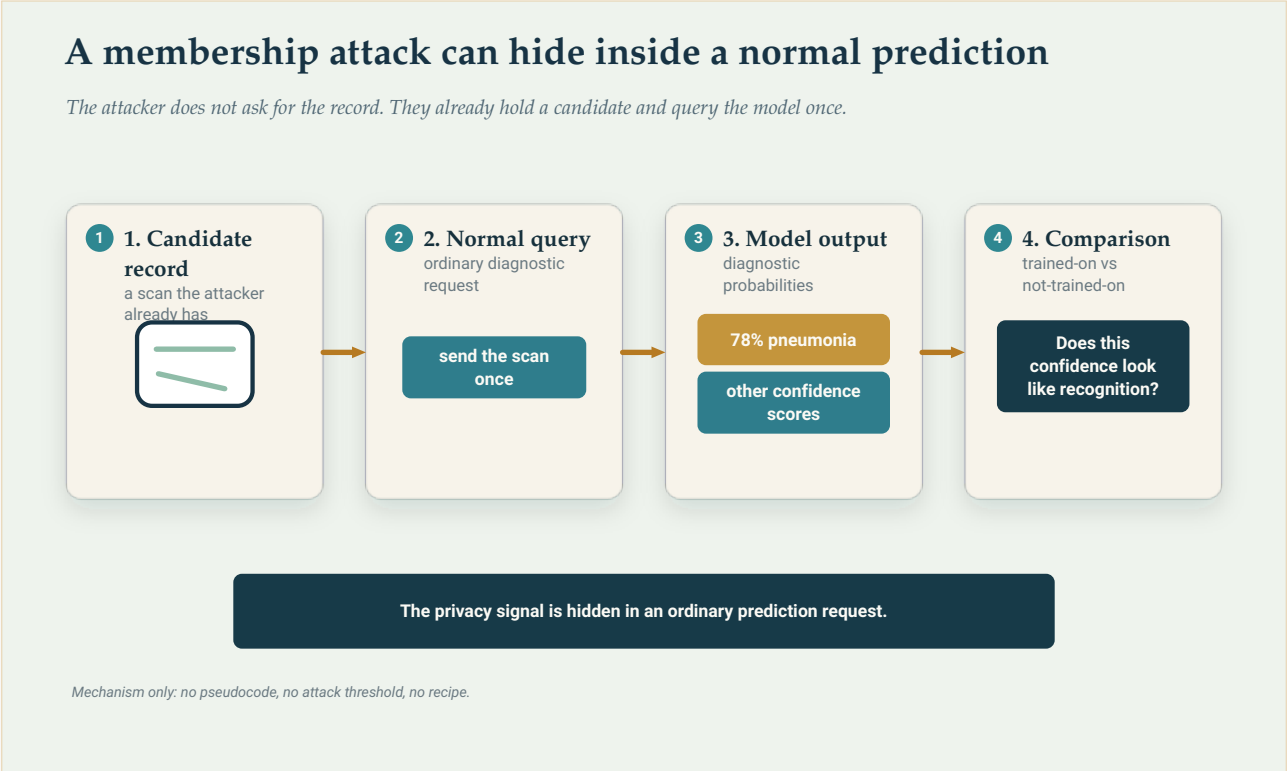
> **Што всушност е membership inference attack**
>
> Истекувањето тука е посуптилно од „моделот го исфрла твоето досие“. Се работи за *членство* — самиот факт дека твоите податоци биле во тренинг-сетот.
> Да почнеме од тоа што нормално прави еден таков модел. Му давате снимка — на пример рендген на граден кош — и тој враќа веројатности: 78% веројатност за пневмонија, заедно со процени за кардиомегалија, едем и консолидација. Нападот го користи *само* тој вообичаен дијагностички одговор. Ништо егзотично.

Слабоста на која се потпира е следнава: моделот обично е малку **посигурен** за точните примери на кои бил обучен отколку за примери што никогаш не ги видел. Замислете ученик кој тајно го видел тестот однапред — на прашањата што веќе ги вежбал одговорите доаѓаат малку пребрзо и со малку премногу сигурност. Да се заклучи од тој сигнал дали конкретен запис бил меѓу примерите што моделот ги видел при обуката е токму значењето на „membership inference“. Нападот, во суштина, изгледа вака. Некој сака да дознае дали снимката на одредено лице била користена за обука. Тој **не** бара моделот да му го врати записот — веќе има кандидат-снимка, оригиналот на лицето или блиска копија. Ја испраќа еднаш како обично дијагностичко барање и ја забележува сигурноста на одговорот. Потоа проверува дали таа сигурност повеќе личи на модел што *бил* обучен со снимката или на модел што *не бил*. Таквата споредба може релативно евтино да ја научи со сопствен заменски „reference model“, обучен да го препознае карактеристичниот образец на преголема сигурност, на обичен компјутер без специјален хардвер. Ако вистинскиот модел е невообичаено сигурен за конкретната снимка — како да ја препознава — тоа е силен сигнал дека снимката била во тренинг-податоците.

Непријатниот дел е што самото барање воопшто не изгледа како напад: тоа е истиот тип дијагностичко барање што би го направил клиничар, а сигналот за приватност е скриен во обично предвидување. Токму затоа ризикот е реален — иако засега е демонстриран под јасно наведени лабораториски претпоставки, а не забележан како напад во реална употреба.

Зошто е важно членството ако самиот запис никогаш не истече? Затоа што *членството е информација*. Ако моделот бил обучен на пациенти што примале конкретна имунотерапија за рак, потврдувањето дека вашиот запис е внатре открива дека веројатно сте го имале тој рак — токму тип информација што осигурител или работодавач не би смеел да може да ја заклучи. Содржината останува затворена; излегува фактот дека припаѓате на групата.

Авторите јасно ги наведуваат условите — пристап до обичните предвидувања на моделот, кандидат-запис и сопствен reference model на напаѓачот — не како упатство за напад, туку за ризикот да може прецизно да се анализира наместо само да се споменува нејасно.



За нападот не е потребен посебен API за приватност. Обично дијагностичко барање може да открие сигнал за членство ако моделот е невообичаено сигурен за запис што веќе го видел.

Што откриле

Три резултати, секој поостар од претходниот.

Просеците ги кријат изложените пациенти. Кога се мери агрегатно — како што обично се прави — многу од овие модели изгледаат охрабрувачки приватни: за повеќето пациенти нападот не е подобар од случаен погодок. Анализата пациент по пациент покажува друга слика. Како што Knolle објаснува во соопштението за студијата, претходните процени го мереле само просечниот ризик за сите пациенти, додека оваа студија за првпат го разгледува ризикот на ниво на поединечен пациент — и сликата е многу поинаква. Кај некои лица нападот успева **речиси совршено**: авторите го мерат тоа како attack AUC од **0,95 или повеќе** (0,5 е случајно погодување, 1,0 е совршено), иако просекот за целиот dataset може да изгледа како случајност.

Просекот може да остане близу до случајно погодување додека мал „опашест“ дел од записите е многу полесно препознатлив. Поентата на трудот е дека приватноста мора да се проверува на ниво на пациент, а не само агрегатно.

Изложеноста не е еднаква — и најмногу ги погодува недоволно застапените. Пациентите најранливи на речиси совршен напад систематски биле од групи **недоволно застапени во податоците**: етнички малцинства, ретки фенотипови на болести и невообичаени карактеристики на снимки. Во dataset со електронски здравствени записи, црните пациенти се појавувале **31% почесто** отколку што би се очекувало меѓу најранливите записи; во dataset со мамографија, снимките означени како сомнителни за малигнитет биле прекумерно застапени во таа крајно ризична група за впечатливи **+1.179%**. (Овие две бројки се примери, не целата слика — трудот го наоѓа истиот образец и кај други групи — и станува збор за *релативна* прекумерна застапеност во малата опашка со екстремен ризик: колку почесто тие пациенти се појавуваат меѓу најизложените записи, а не колкав дел од сите црни пациенти или од сите сомнителни мамографии е изложен.) Моделот има помалку слични примери со кои би ги „стопил“ овие пациенти во мнозинството, па нивните записи повеќе се издвојуваат — а токму тоа издвојување го фаќа membership attack. Неуспехот на приватноста не е случајно распределен; се концентрира кај луѓето што веќе се на маргините.

Поголемите модели го влошуваат проблемот. Бројот пациенти изложени на речиси совршени напади нагло растел со капацитетот на моделот — во еден дерматолошки dataset, уделот се искачил од практично нула кај најмалиот модел до околу **еден од десет** кај најголемиот. Додека медицинските AI-модели стануваат поголеми и поспособни — насоката во која се движи целото поле — овој конкретен ризик,

предупредуваат авторите, станува посериозен, а не помал.

Rückert го сумира остро: „Ова не е прифатлив ризик. Здравствените податоци се исклучително чувствителни.“

Зошто „безбедно во просек“ е погрешниот тест

Лесно е нискиот просечен ризик да се прочита како потврда дека сè е во ред. Главната лекција на трудот е дека просекот тука не е само непрецизен — туку го мери погрешното нешто.

Гаранцијата за приватност има смисла само ако важи и за лицето со најголем ризик, не само за „просечниот“ пациент. Модел во кој 999 од 1.000 пациенти не можат да се препознаат, а еден може речиси совршено да се идентификува, не е „99,9% приватен“ во смисла што му значи на тој еден човек. А ако тој човек предвидливо е пациент со ретка болест или припадник на етничко малцинство, метриката не е само нецелосна — таа тивко создава нееднаква заштита. Ја мери безбедноста на мнозинството и ја именува како безбедност за сите.

Токму тоа е промената што ја наметнува трудот: од *колку е приватен моделот во просек?* кон *кој пациент е најизложен, и кој е тој човек?* Тоа се различни прашања, а само второто ја тестира приватноста како ветување кон поединецот.

Што ова *не* докажува — ниту тврди

- Не вели дека медицинскиот AI треба да се напушти. Рамката на авторите е намалување на ризикот, не повлекување: измери го и поправи го ризикот, не престанувај да градиш.
- Не значи дека секој медицински модел „протекува“, ниту дека некој конкретен модел во употреба бил нападнат. Трудот покажува дека *ризикот постои и е нееднакво распределен*, и дека стандардните агрегатни метрики го промашуваат.
- Не значи дека вашите записи веќе се изложени. Нападот бара конкретни услови — пристап до моделот, кандидат-запис и сопствена инфраструктура на напаѓачот — не е нешто што се случува случајно.
- Не покажува дека истекува *содржината* на пациентскиот запис. Она што се открива е членството — фактот дека записот бил вклучен — што е опасно од друга причина, а не затоа што досието се извлекува од моделот.
- Не ги сведува разликите на една единствена впечатлива бројка. Силниот и повторлив резултат е *образецот* — агрегатните метрики го потценуваат индивидуалниот ризик, а најголемиот товар паѓа врз недоволно застапените пациенти — и тој се гледа низ datasets и стотици модели.

Колку се убедливи доказите?

Ова е емпириски и методолошки резултат, и е робустен. Образецот — агрегатните метрики систематски го потценуваат ризикот по пациент, преостанатиот ризик се концентрира кај недоволно застапените групи, а проблемот се влошува со капацитетот на моделот — се повторува низ **седум datasets со различни типови податоци** и голем број модели во секој од нив. Токму таа ширина го прави тврдењето за мерењето убедливо, наместо артефакт на еден dataset.

Има две важни ограничувања. Прво, ова е демонстрација на *ризик*, измерен со напади што самите автори ги спровеле; покажува колку добро способен напаѓач *би можел* да успее под наведените претпоставки, а не колку често вакви напади се случуваат во реалниот свет. Второ, впечатливите бројки — речиси совршен успех кај некои пациенти — намерно ги опишуваат најизложените поединци; тоа е и суштината на анализата. Треба да се читаат како „опашката на распределбата е многу потешка отколку што сугерира просекот“, а не како „повеќето пациенти се изложени“.

Соодветниот став не е ниту паника ниту отфрлање: ова е внимателна студија на повеќе datasets што покажува дека стандардниот начин на сертифицирање приватност кај медицински AI е слеп за сопствените најлоши случаи — и дека тие случаи ги погодуваат пациентите што имаат најмалку простор за дополнителен ризик.

Зошто е важно

Две работи го прават ова повеќе од техничка фуснота.

Прво, го менува она што смееме да го наречеме „зачувување на приватноста“. Ако модел се објави со ниска просечна вредност за ризик, таа бројка навистина може да биде ниска, а сепак да крие мала група пациенти што membership attack може речиси совршено да ги препознае. Практичното барање на трудот е конкретно: пред објавување да се проценува ризикот за приватност **за секој пациент**, да се контролира кој има пристап до моделите во употреба и да се користат техники како **диференцијална приватност** — мала, внимателно калибрирана количина шум додадена за време на обуката, која ги отежнува membership нападите, со реална и контролирана цена во корисноста на моделот (privacy–utility trade-off постои и не е бесплатен). „Го проверивме просекот“ не треба повеќе да значи „ја проверивме приватноста“.

Второ, трудот ја врзува приватноста со правичноста. Истите групи што се недоволно застапени во медицинските податоци — и затоа веќе добиваат полоша услуга од медицински AI — излегуваат како групи чија приватност истиот AI ја штити најслабо. Поле што напорно работи на првата нееднаквост не може втората да ја третира како туѓ проблем. Станува збор за истите луѓе.

Утешната приказна за приватноста во медицински AI — *ја измеривме, просекот е низок, значи сме во ред* — е токму она што овој труд го разложува. Не за да ги исплаши луѓето од технологијата, туку за стандардот да се помести таму каде што припаѓа: ветување што можете да тврдите дека го исполнувате само ако сте го провериле за лицето што најлесно може да биде повредено.

Кратко резиме

Membership inference нападите се обидуваат да утврдат дали записот на конкретно лице бил во податоците за обука на AI-модел — а бидејќи самото членство може да открие чувствителна информација (на пример дека сте имале одредена болест), тоа е вистинско нарушување на приватноста и кога самиот запис никогаш не се појавува. Претходните работи најчесто мереле колку често нападот успева *во просек* за цел dataset. Оваа студија го мери ризикот *по пациент*, низ седум медицински datasets (снимки, ECG, електронски записи) и многу модели во секој, и наоѓа дека агрегатните метрики сериозно го потценуваат ризикот: некои поединци може речиси совршено да се идентификуваат (attack AUC $\geq$ 0,95) иако просекот за dataset изгледа безбедно; најранливи систематски се луѓето од недоволно застапени групи (етнички малцинства, ретки болести, невообичаени снимки); а проблемот расте со големината на моделот. Авторите не предлагаат напуштање на медицинскиот AI — предлагаат приватноста да се мери на индивидуално ниво, пристапот до моделите да се контролира и да се користи диференцијална приватност. Непријатната суштина е: „приватно во просек“ не е гаранција за приватност, а најчесто не ги штити токму пациентите што и инаку се најслабо заштитени.

Проверка без претерување

Што покажува трудот: Низ седум медицински datasets и многу модели во секој од нив, ризикот од membership inference за одредени пациенти е многу повисок отколку што сугерираат агрегатните метрики — до речиси совршен успех на нападот (AUC $\geq$ 0,95) — а тој преостанат ризик непропорционално паѓа врз недоволно застапени групи и расте со капацитетот на моделот.

Што е веројатно, но не е докажано: Дека напаѓачи во реалниот свет веќе го користат ова против клинички модели во употреба; студијата ја демонстрира *можноста и нејзината распределба* под наведени претпоставки, а не зачестеноста на напади во практиката.

Што не покажува: Дека медицинскиот AI треба да се напушти; дека секој модел протекува или дека некој конкретен модел во употреба бил пробиен; дека е изложена *содржината* на пациентските записи наместо само членството; една единствена

бројка што ја опишува целата нееднаквост — робустниот резултат е образецот, не една бројка.

Главни ограничувања: Го мери најлошиот ризик преку напади изведени од самите автори, а не преку забележани инциденти; драматичните бројки намерно ги опишуваат најизложените поединци; точните разлики меѓу групите се прикажани како конзистентен образец низ datasets, а не сведени на еден број.

Колкава доверба треба да има општ читател? Висока доверба дека агрегатните метрики го потценуваат индивидуалниот ризик, дека преостанатиот ризик се концентрира кај недоволно застапените пациенти и дека се влошува со големината на моделот — тоа е покажано низ многу datasets и модели. Умерена доверба за тоа колку често вакви напади се случуваат во реалниот свет, бидејќи студијата тоа не го мери. Безбедното читање не е „медицинскиот AI ги протекува твоите податоци“, ниту „приватноста е решена“, туку: „стандардната проверка за приватност е слепа за сопствените најлоши случаи, а тие случаи паѓаат врз најранливите пациенти — затоа проверката мора да се смени“.

Извори

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