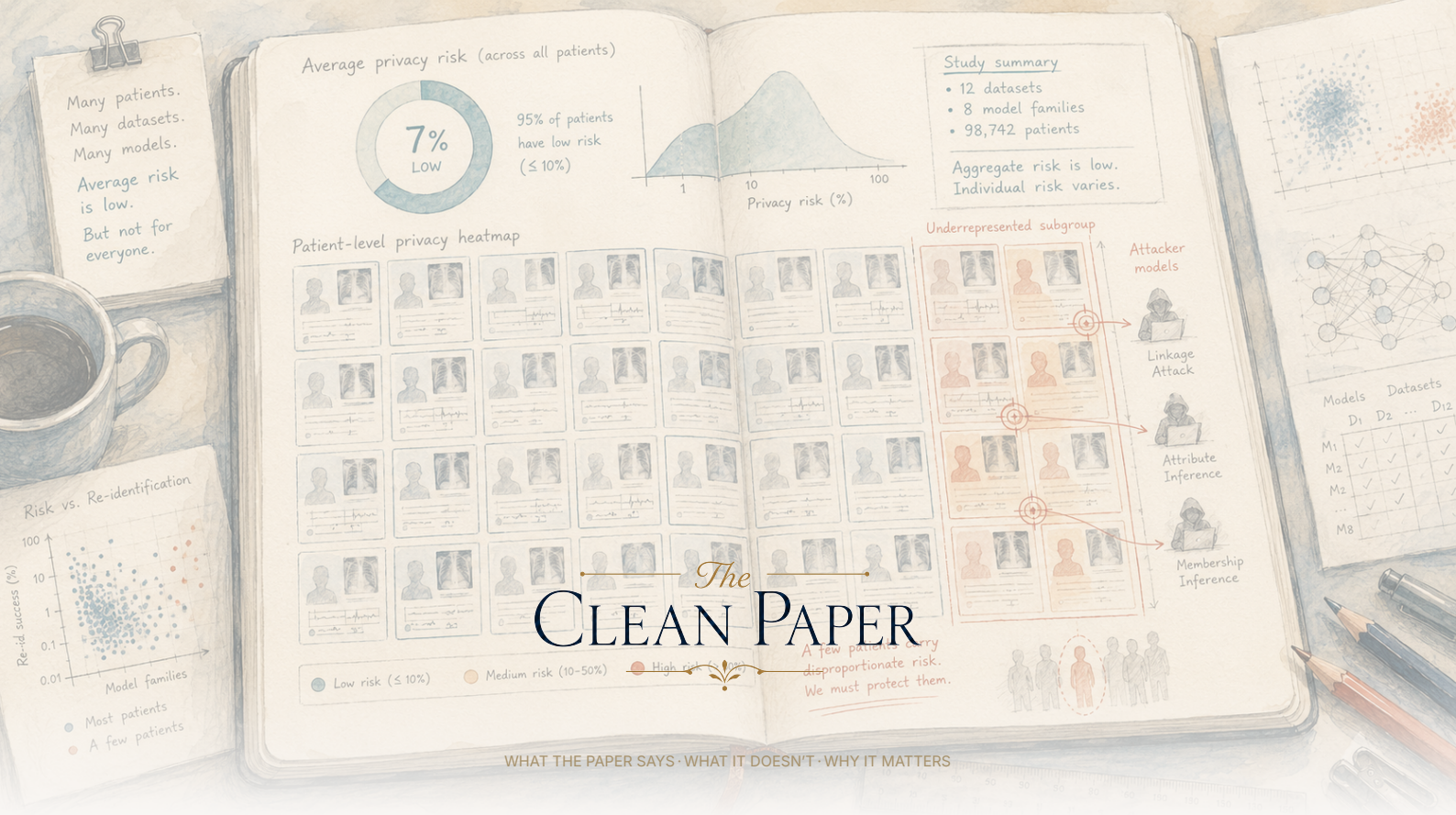


Медицински AI може да е поверителен средно и въпреки това да излага конкретни пациенти — най-вече недостатъчно представените

Медицински AI модел, наречен „privacy-preserving“, обикновено се опира на едно средно число. Това изследване твърди, че това е грешният тест. Изучавайки membership inference attacks — които разкриват дали записът на конкретен човек е бил в тренировъчните данни и така могат да издадат, че той е имал дадено заболяване — авторите измерват риска по пациент, а не агрегирано, в седем медицински набора данни (образна диагностика, ECG, електронни досиета) и множество модели във всеки. Модели, които изглеждат безопасни средно, все пак могат да позволят почти перфектно идентифициране на конкретни хора (attack AUC ≥ 0,95); изложените систематично са от недостатъчно представени групи (етнически малцинства, редки заболявания, необичайни образни характеристики); и проблемът се влошава с увеличаването на моделите. Авторите не казват да се изостави медицинският AI — казват да се измерва поверителността по пациент, да се контролира достъпът до моделите и да се използва differential privacy. Неприятното ядро: „поверително средно“ не е гаранция за поверителност и се проваля точно за пациентите, които вече са най-слабо защитени.

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

Гаранцията за поверителност е обещание към човек, не към средна стойност

Когато медицински AI модел бъде наречен „запазващ поверителността“, твърдението обикновено се опира на едно число: сред всички пациенти, чиито данни са обучавали модела, средната вероятност да се разбере дали даден човек е бил в тренировъчния набор е ниска. Това звучи успокояващо. Тихата и неудобна идея на тази статия е, че това е грешното число.

Поверителността не е средна стойност. Тя е обещание към всеки отделен човек — че присъствието му в тренировъчните данни няма по-късно да го изложи. А средната стойност може да спази обещанието за почти всички и напълно да го наруши за малцина. Изследователите измерват риска един пациент по един, с резолюция, която не е използвана досега, и намират точно това: модели, които изглеждат безопасни агрегирано, могат почти перфектно да издават принадлежността на конкретни хора — и непропорционално често това са хората, които и без това са най-слабо защитени.

Какво направиха авторите

Екипът — ръководен от Moritz Knolle, с Daniel Rückert (и двамата от Technical University of Munich) и Georg Kaissis (Hasso Plattner Institute), заедно с колеги от Imperial College London — взема добре позната заплаха за поверителността, наречена **атака за установяване на принадлежност (membership inference attack)**, и променя въпроса. Вместо „средно колко често тази атака успява в целия набор данни?“, те питат „за *този конкретен пациент* колко е изложен?“

Те провеждат анализа по пациенти върху **седем утвърдени медицински набора данни**, обхващащи много различни типове — медицински изображения, електрокардиограми и електронни здравни досиета — и по 200 модела във всеки. За всеки пациент във всеки модел оценяват колко уверено атакуващ може да определи дали записът на човека е бил част от тренировъчните данни, след което разбиват резултатите по групи: статус на заболяване, самоопределена раса, осигуряване, пол и протокол на образната диагностика.

Какво всъщност е атака за установяване на принадлежност

Течът тук е по-фин от „моделът изплюва досието ви“. Става дума за *принадлежността* — самия факт, че данните ви са били в тренировъчния набор.

Да започнем с обикновената работа на такъв модел. Давате му скен — например рентгенография на гръден кош — и той връща вероятности: 78% вероятност за пневмония, плюс оценки за кардиомегалия, оток, консолидация. Този ежедневен

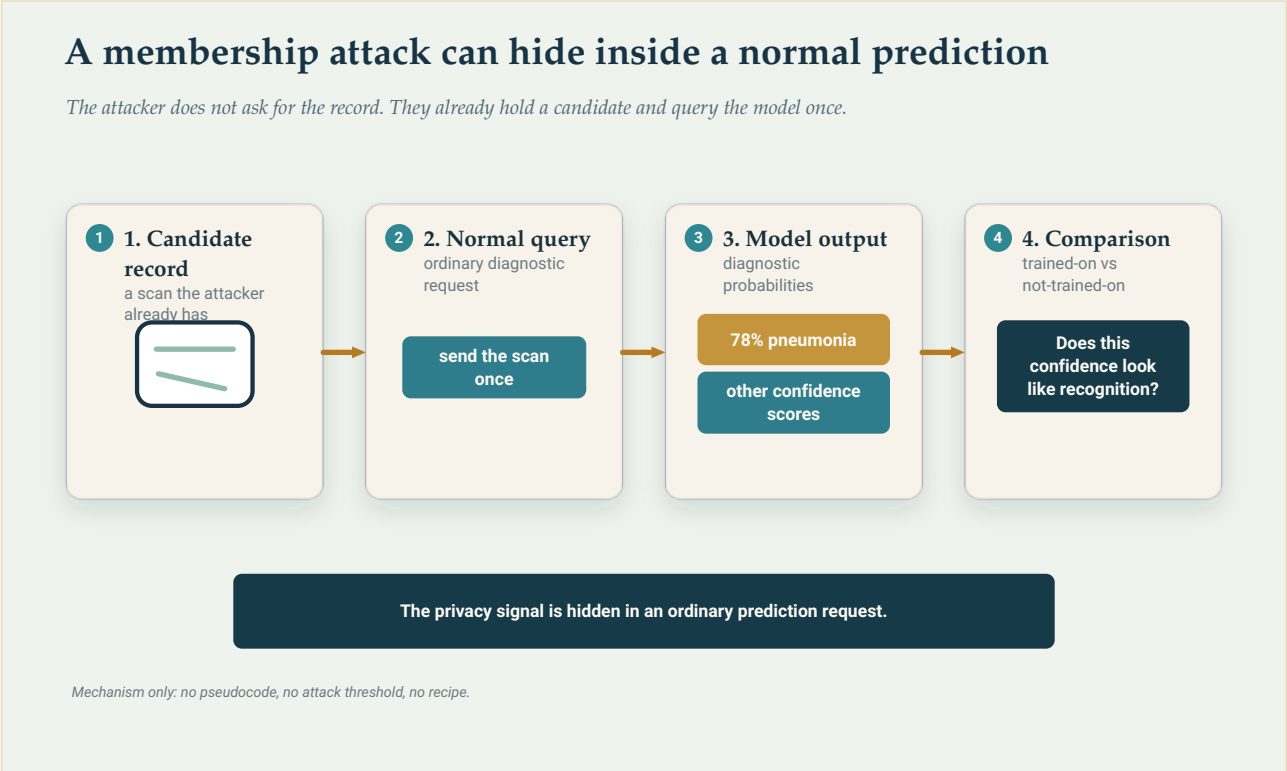
диагностичен отговор е *единственото*, което атаката използва. Нищо екзотично. Слабостта е следната: моделът обикновено е малко **по-уверен** за точно онези примери, върху които е обучаван, отколкото за такива, които никога не е виждал. Представете си ученик, който тайно е виждал изпита предварително — на въпросите, които е упражнявал, отговорите идват прекалено бързо и прекалено уверено. Да се определи от тази следа дали конкретен запис е бил сред примерите, които моделът е изучавал, е това, което означава „установяване на принадлежност“.

Атаката стъпка по стъпка изглежда така. Някой иска да знае дали скенът на конкретен човек е използван за обучение на модела. Той **не** иска досието от модела — вече има кандидат скен (самия скен на човека или близко копие). Изпраща го веднъж като обикновена диагностична заявка и записва колко уверен е отговорът. После проверява дали тази увереност прилича повече на модел, който *е* тренирал върху скена, или на такъв, който *не е* — сравнение, което може да направи евтино, като обучи собствен заместител („референтен модел“), за да научи как изглежда характерната свръхувереност, на обикновен компютър и без специален хардуер. Ако истинският модел е необичайно сигурен за този скен — сякаш го разпознава — това е силно доказателство, че скенът е бил в тренировъчните данни.

Неприятното е, че нищо в заявката не изглежда като атака: тя е същата диагностична заявка, която би направил клиницист, а сигналът за поверителност е скрит в обикновена прогноза. Точно затова рискът е конкретен — макар засега да е демонстриран при заявени лабораторни допускания, не уловен в реална атака.

Защо принадлежността има значение, ако самото досие никога не изтича? Защото *принадлежността е факт*. Ако моделът е обучен върху пациенти, получавали конкретна имунотерапия за рак, потвърждението, че вашият запис е в него, разкрива, че вероятно сте имали този рак — точно вид информация, която застраховател или работодател не бива да може да извлече. Съдържанието остава затворено; изтича фактът, че принадлежите към групата.

Авторите излагат тези допускания ясно — достъп до обикновените прогнози на модела, кандидат запис и собствен референтен модел на атакуващия — не като инструкции за атака, а за да може заплахата да бъде анализирана, вместо замахната с ръка.



Атаката не се нуждае от специален API за поверителност. Нормална диагностична заявка може да издаде сигнал за принадлежност, ако моделът е необичайно уверен върху запис, който вече е виждал.

Какво откриха

Три резултата, всеки по-остър от предишния.

Средните стойности скриват изложените. Измерени агрегирано — обичайният начин — много от тези модели изглеждат успокояващо поверителни: за повечето пациенти атаката не е по-добра от случайността. Прегледът по пациенти разказва друга история. Както Knolle казва в съобщението за изследването, предишните оценки „са измервали само средния риск за всички пациенти. Ние за първи път разгледахме риска на ниво отделен пациент — и картината е много различна.“ За някои хора атаката успява **почти перфектно** — авторите го измерват като AUC на атаката **0,95 или повече** (0,5 е хвърляне на монета, 1,0 е безгрешно) — дори когато средното за целия набор изглежда не по-добре от случайността.

Средната стойност може да стои близо до случайността, докато малка опашка от записи остава много по-изложена. Идеята на статията е, че поверителността трябва да се проверява на ниво пациент, не само агрегирано.

Излагането е неравномерно — и пада върху недостатъчно представените. Пациентите, най-уязвими на почти перфектна атака, систематично са от групи, **недостатъчно представени в данните**: етнически малцинства, фенотипове на редки заболявания, необичайни образни характеристики. В набора с електронни досиета чернокожи пациенти се появяват **31% по-често** от очакваното сред най-уязвимите записи; в набора с мамографии скенове, обозначени като подозрителни за злокачественост, са свръхпредставени в опасната зона с впечатляващи **+1 179%**. (Тези две числа са примери, не цялата картина — статията съобщава същия наклон в няколко групи — и са *относителни* свръхпредставяния в малката крайно рискова опашка: колко по-често тези пациенти се появяват сред най-изложените записи, не какъв дял от всички чернокожи или подозрителни мамографски пациенти са изложени.) Моделът има по-малко подобни примери, в които тези пациенти да се „разтворят“, затова записите им изпъкват — а точно изпъкването е това, което атака за установяване на принадлежност улавя. Провалът на поверителността не е случаен; концентрира се върху хората, които вече са на ръба.

По-големите модели влошават проблема. Броят пациенти, изложени на почти перфектни атаки, расте рязко с капацитета на модела — в един дерматологичен набор делът се покачва от практически нула при най-малкия модел до около **един на десет** при най-големия. Докато медицинските AI модели стават по-големи и способни — точно посоката, в която се движи областта — този конкретен риск, предупреждават

авторите, става по-сериозен, не по-малък.

Резюмето на Rückert е директно: „Това не е приемлив риск. Здравните данни са силно чувствителни.“

Защо „безопасно средно“ е грешният тест

Изкушаващо е ниският среден риск да се прочете като чисто удостоверение за безопасност. Основният урок на статията е, че усредняването тук не е просто неточно — то измерва грешното нещо.

Гаранцията за поверителност има смисъл само ако важи за човека с най-голям риск, не за човека по средата. Модел, при който 999 от 1 000 пациенти не могат да бъдат идентифицирани, но един може почти перфектно, не е „99,9% поверителен“ в смисъл, който има значение за този един човек — а ако този човек предсказуемо е пациентът с рядко заболяване или от етническо малцинство, метриката не е просто непълна, а тихо дискриминираща. Тя съобщава безопасност за мнозинството и я нарича безопасност за всички.

Това е промяната, която статията налага: от *колко поверителен е този модел средно?* към *кой е най-изложеният пациент и кой е той?* Това са различни въпроси и само вторият е истински въпрос за поверителността.

Какво *не* доказва това — или твърди

- То **не** казва медицинският AI да бъде изоставен. Рамката на авторите е смекчаване на риска, не отстъпление: измерете и поправете риска, не спирайте да строите.
- То **не** означава, че всеки медицински модел изтича данни или че конкретен внедрен модел е бил атакуван. Показва, че *рискът съществува и е неравномерно разпределен* и че стандартните агрегирани метрики го пропускат.
- То **не** означава, че досиетата ви вече са изложени. Атаката изисква конкретни условия — достъп до модела, кандидат запис и собствена инфраструктура на атакуващия — не е случайна възможност.
- То **не** показва, че *съдържанието* на пациентското досие изтича. Изтича принадлежността — фактът на включване — който е опасен по различна причина, не защото записът се отпечатва навън.
- То **не** свежда неравенствата до едно заглавно число. Силният и възпроизводим резултат е *моделът* — агрегатните метрики подценяват индивидуалния риск и недостатъчно представените пациенти носят най-много от него — в много набори данни и стотици модели.

Колко силни са доказателствата?

Това е емпиричен, методологичен резултат и е стабилен. Моделът — агрегатните метрики за поверителност систематично подценяват риска за отделни пациенти, остатъчният риск се концентрира върху недостатъчно представени групи и се влошава с капацитета на модела — се запазва в **седем набора данни от различни типове** и голям брой модели във всеки. Именно тази ширина прави измервателното твърдение достоверно, а не артефакт на един набор.

Две честни уговорки. Първо, това е демонстрация на *риск*, измерен чрез атаките, изградени от самите автори; тя характеризира колко добре способен атакуващ *би могъл* да се справи при заявените допускания, не колко често реални атаки се случват. Второ, впечатляващите числа — почти перфектен успех за някои пациенти — описват най-зле защитените хора *по замисъл*; точно това е смисълът и трябва да се чете като „опашката е много по-тежка, отколкото средната предполага“, не като „повечето пациенти са изложени“.

Подходящата позиция не е нито тревога, нито пренебрежение: внимателно многонаборно изследване показва, че стандартният начин да удостоверяваме поверителността на медицински AI е сляп за собствените си най-лоши случаи — и че тези случаи падат върху пациентите с най-малко запас за грешка.

Защо това има значение

Две неща правят резултата повече от техническа бележка.

Първо, той променя какво трябва да е позволено да означава „съхраняващ поверителността“. Ако модел бъде публикуван със средна оценка за поверителност, тя може действително да е ниска и все пак да крие подгрупа пациенти, почти перфектно идентифицируеми от атака за установяване на принадлежност. Практическото искане на статията е конкретно: оценявайте риска за поверителността **по пациент** преди пускане, контролирайте кой има достъп до внедрените модели и използвайте техники като **диференциална поверителност** — малък, внимателно калибриран шум, добавен по време на обучението, който притъпява атаки за установяване на принадлежност срещу реална и управляема цена в полезността на модела (компромисът поверителност–полезност е изричен, не безплатен). „Проверихме средната стойност“ трябва да престане да се брои като „проверихме“.

Второ, резултатът сплита поверителността със справедливостта. Същите групи, които са недостатъчно представени в медицинските данни — и затова вече са обслужвани по-зле от медицински AI — се оказват групите, чиято поверителност същият AI защитава най-слабо. Област, която усилено работи по първото неравенство, не може да третира второто като чужд проблем. Това са същите хора.

Успокояващата история за поверителността на медицинския AI — *измерихме я, средната стойност е ниска, всичко е наред* — е точно историята, която тази статия разглобява. Не за да изплаши хората от технологията, а за да премести стандарта там, където му е мястото: обещание, което можете да твърдите, че спазвате, само ако сте го проверили за човека, който най-вероятно ще пострада.

Кратко и ясно

Атаките за установяване на принадлежност се опитват да определят дали записът на конкретен човек е бил в тренировъчните данни на AI модел — и понеже самата принадлежност може да разкрива нещо (например, че сте имали определено заболяване), това е истински теч на поверителност, дори когато самият запис никога не се появява. По-ранните работи измерват колко често такива атаки успяват *средно* в набора. Това изследване измерва риска *по пациент* в седем медицински набора данни (образна диагностика, ECG, електронни досиета) и много модели във всеки и намира, че агрегатните метрики силно подценяват риска: някои хора могат да бъдат идентифицирани почти перфектно (AUC на атаката $\geq$ 0,95), дори когато средната стойност изглежда безопасна; най-уязвими систематично са хората от недостатъчно представени групи (етнически малцинства, редки заболявания, необичайни изображения); а проблемът расте с размера на модела. Авторите не призовават да се изостави медицинският AI — призовават да се измерва поверителността индивидуално, да се контролира достъпът до моделите и да се използва диференциална поверителност. Изводът: „поверително средно“ не е гаранция за поверителност, а хората, които този тест подвежда, обикновено са вече най-слабо защитените.

Проверка без излишен шум

Какво показва статията: В седем медицински набора данни и множество модели във всеки рискът от установяване на принадлежност за отделни пациенти е много по-висок за някои хора, отколкото агрегатните метрики предполагат — до почти перфектен успех на атаката (AUC $\geq$ 0,95) — а остатъчният риск пада непропорционално върху недостатъчно представени групи и расте с капацитета на модела.

Какво е правдоподобно, но не е доказано: Че реални атакуващи вече използват това срещу внедрени клинични модели; изследването демонстрира *способността и разпределението ѝ* при заявени допускания, не честотата на атаки в реалността.

Какво не показва: Че медицинският AI трябва да бъде изоставен; че всеки модел изтича данни или че конкретен внедрен модел е пробит; че *съдържанието* на пациентските досиета, а не принадлежността, е изложено; едно-единствено количествено число за неравенството — стабилният резултат е моделът, не една стойност.

Основни ограничения: Измерва най-лошия риск чрез атаките на самите автори, не наблюдавани инциденти; драматичните числа по дизайн описват най-изложените хора; точните групови неравенства се показват като последователен модел между наборите, а не се свеждат до едно число.

Колко уверен трябва да бъде общият читател? Висока увереност, че агрегатните метрики за поверителност подценяват индивидуалния риск, че остатъчният риск се концентрира върху недостатъчно представени пациенти и че това се влошава с размера на модела — демонстрирано е в много набори и модели. Умерена увереност за реалната честота на такива атаки, която това изследване не измерва. Безопасният прочит не е „медицинският AI изтича данните ви“ и не е „поверителността е решена“, а „стандартната проверка е сляпа за собствените си най-лоши случаи и тези случаи падат върху най-уязвимите пациенти — затова проверката трябва да се промени“.

Източници

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