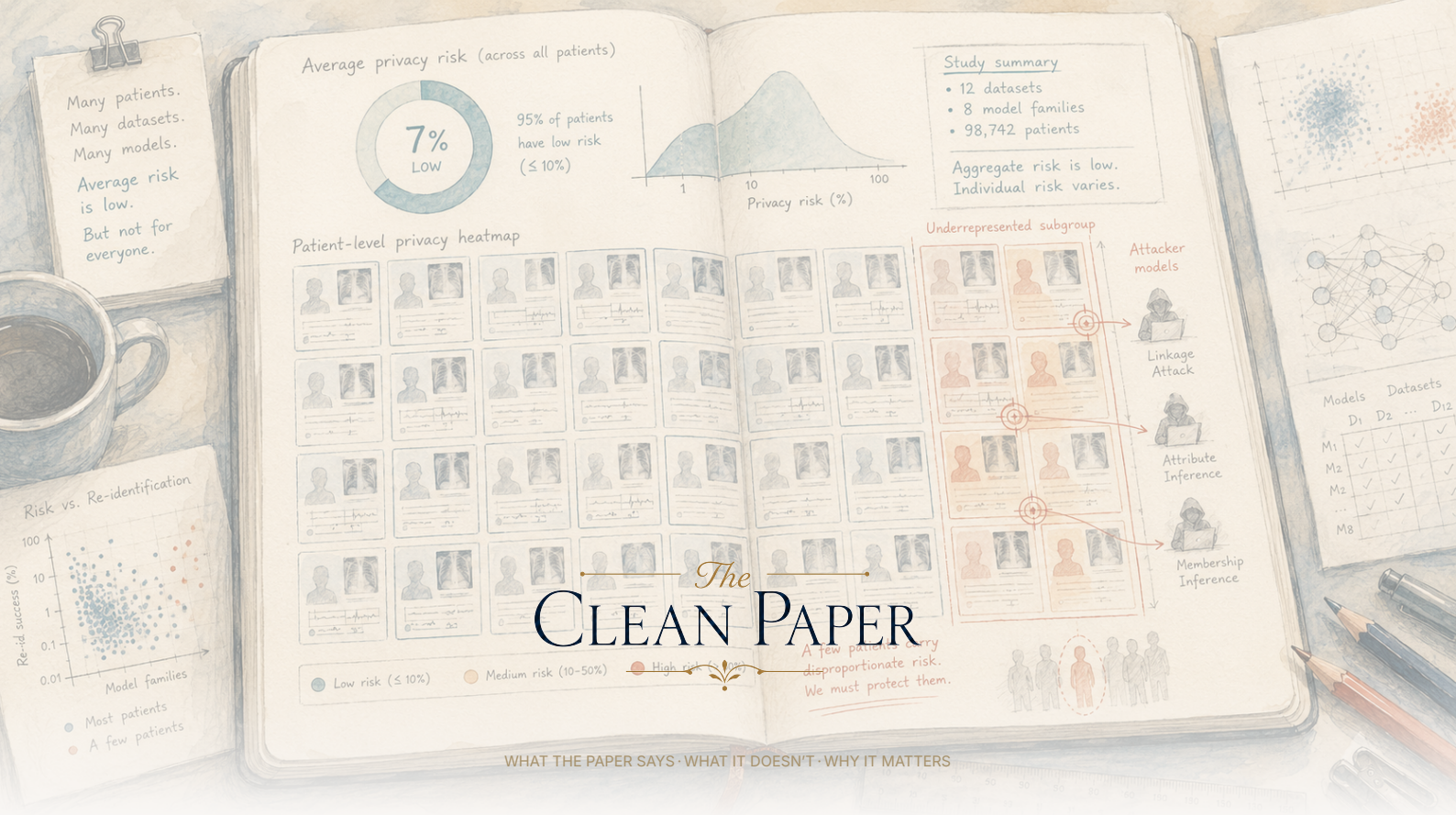


Медичний ШІ може бути приватним у середньому й водночас викривати окремих пацієнтів — найбільше недопредставлених

Коли медичну модель ШІ називають такою, що «зберігає приватність», це зазвичай спирається на один середній показник. Це дослідження показує, що така перевірка неправильна. Вивчаючи атаки на визначення належності до навчальної вибірки — вони дозволяють з'ясувати, чи був запис конкретної людини в навчальних даних моделі й тим самим можуть розкрити, що людина мала певну хворобу, — автори вимірювали ризик не в сукупності, а для кожного пацієнта окремо, на семи медичних датасетах (зображення, ЕКГ, електронні записи) і багатьох моделях. Картина така: моделі, безпечні за середнім показником, іноді дозволяють майже безпомилково ідентифікувати конкретних людей (AUC атаки ≥ 0,95); серед викритих систематично переважають недопредставлені групи — етнічні меншини, рідкісні хвороби, незвичайні зображення; і зі збільшенням моделей проблема посилюється. Автори не закликають відмовитися від медичного ШІ — вони пропонують вимірювати приватність для кожного пацієнта, контролювати доступ до моделей і застосовувати диференційну приватність. Незручна суть: «приватна в середньому» — не гарантія приватності, і найчастіше вона підводить тих пацієнтів, які вже захищені найменше.

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

Гарантія приватності — це обіцянка людині, а не середньому значенню

Коли про медичну модель ШІ кажуть, що вона «зберігає приватність», зазвичай це твердження спирається на одне число: для всіх пацієнтів, на даних яких навчали модель, середня ймовірність визначити, чи входили дані конкретної людини до навчального набору, є низькою. Це звучить заспокійливо. Тихий і незручний висновок цієї роботи полягає в тому, що це просто не те число.

Приватність — не середнє значення. Це обіцянка кожній окремій людині: участь її даних у навчальному наборі не повинна згодом її викрити. А середній показник може виконувати цю обіцянку майже для всіх і водночас повністю порушувати її для кількох людей. Дослідники виміряли ризик приватності для кожного пацієнта окремо, з роздільною здатністю, якої раніше не використовували, й побачили саме це: моделі, які в сукупності виглядають безпечними, можуть майже безпомилково видавати факт участі конкретних людей у навчальних даних — і непропорційно часто це саме ті люди, які й без того найменш захищені.

Що зробили автори

Команда на чолі з Моріцем Кнолле, разом із Даніелем Рюккертом (обидва — Technical University of Munich) і Георгом Кайссісом (Hasso Plattner Institute), а також колегами з Imperial College London взяла добре відому загрозу приватності — **атаку на визначення належності до навчальної вибірки (membership inference attack)** — і змінила запитання. Замість «як часто ця атака в середньому успішна по всьому датасету?» вони запитали: «наскільки вразливий *цей конкретний пацієнт*?»

Такий аналіз на рівні окремого пацієнта вони провели на **семи усталених медичних наборах даних**, що охоплювали дуже різні типи інформації — медичні зображення, електрокардіограми та електронні медичні записи, — і для кожного з них дослідили по 200 моделей. Для кожного пацієнта в кожній моделі вони оцінювали, наскільки впевнено зловмисник міг би визначити, чи була картка цієї людини частиною навчальних даних, а потім розкладали результати за групами: статусом хвороби, самоідентифікованою расовою належністю, страхуванням, статтю та протоколом візуалізації.

> **Що насправді таке атака на визначення належності до навчальної вибірки**
>
> Витік тут тонший, ніж «модель видає вашу медичну картку». Йдеться про *належність* — сам факт того, що ваші дані були в навчальному наборі.
> Почнімо з того, що така модель робить у звичайному режимі. Ви подаєте їй зображення — скажімо, рентген грудної клітки — й отримуєте ймовірності: 78%

імовірності пневмонії, а також оцінки для кардіомегалії, набряку, консолідації. Саме цю повсякденну діагностичну відповідь і використовує атака. Нічого екзотичного.

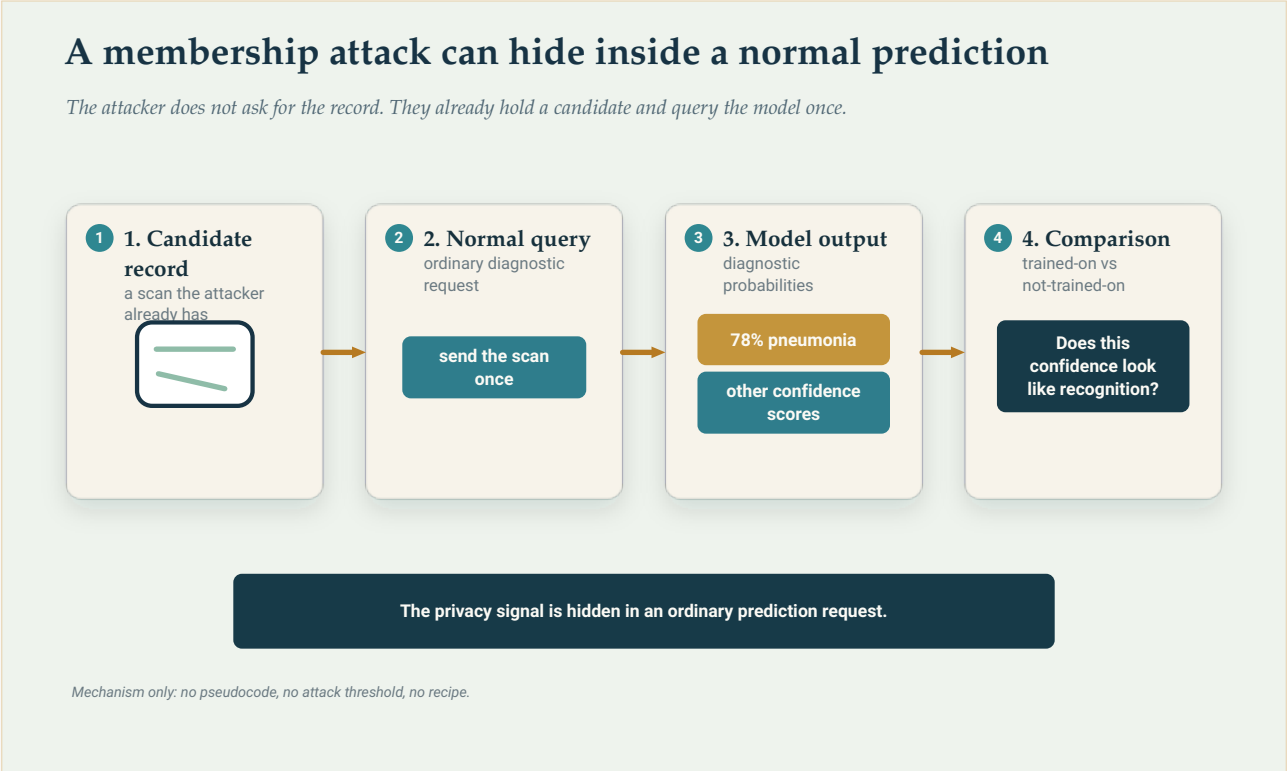
Слабке місце таке: модель зазвичай трохи **впевненіша** саме на тих прикладах, на яких її навчали, ніж на тих, яких вона ніколи не бачила. Уявіть студента, який потай бачив екзаменаційні питання заздалегідь: на вже відпрацьовані запитання відповіді приходять трохи швидше й трохи впевненіше. Визначити за цією ознакою, чи був конкретний запис серед прикладів, які «вивчала» модель, — це і є membership inference.

Отже, атака виглядає так. Хтось хоче дізнатися, чи використали скан конкретної людини для навчання моделі. Він **не** просить модель віддати запис — у нього вже є скан-кандидат, власний скан цієї людини або дуже близька копія. Він один раз подає його як звичайний діагностичний запит і дивиться, наскільки впевнена відповідь. Потім порівнює цю впевненість із тим, як поводиться модель, що *навчалася* на такому скані, і модель, що *не навчалася* на ньому. Таке порівняння можна недорого побудувати, навчивши власну допоміжну «референсну модель», щоб вона показала характерну надмірну впевненість; для цього достатньо звичайного комп'ютера без спеціального обладнання. Якщо справжня модель незвично впевнена саме на цьому скані — ніби впізнає його, — це сильний сигнал, що скан входив до навчальних даних.

Найнеприємніше те, що сам запит зовсім не виглядає як атака: це той самий діагностичний запит, який міг би зробити клініцист, а сигнал про приватність захований усередині звичайного прогнозу. Саме тому ризик конкретний — хоча поки що його демонстрували за чітко визначених лабораторних припущень, а не ловили в реальних атаках.

Навіщо взагалі важлива належність, якщо сам запис ніколи не витікає? Тому що *належність теж є фактом*. Якщо модель навчали на пацієнтах, які отримували певну імунотерапію раку, то підтвердження того, що ваш запис входив до цього набору, може розкрити, що ви, ймовірно, мали цей рак — саме той тип інформації, який страховик або роботодавець не повинен мати змоги вивести. Вміст запису лишається закритим; назовні виходить факт належності до групи.

Автори прямо описують припущення атаки — доступ до звичайних прогнозів моделі, наявність запису-кандидата та власної референсної моделі зловмисника — не як інструкцію, а щоб загрозу можна було аналізувати конкретно, а не відмахуватися від неї.



Для атаки не потрібен спеціальний API приватності. Звичайний діагностичний запит може нести сигнал про належність до навчальної вибірки, якщо модель незвично впевнена на записі, який уже бачила.

Original Aurora diagram — The Clean Paper CC BY 4.0

Що вони знайшли

Три результати, кожен гостріший за попередній.

Середні значення приховують найбільш уразливих. Якщо вимірювати в сукупності — як це зазвичай роблять, — багато моделей виглядають досить приватними: для більшості пацієнтів атака не краща за випадкове вгадування. Але на рівні окремої людини картина інша. Як Кнолле сказав у повідомленні про дослідження, попередні оцінки «завжди вимірювали лише середній ризик для всіх пацієнтів. Ми вперше дослідили ризик на рівні окремих пацієнтів — і картина виявилася зовсім іншою». Для деяких людей атака працює **майже ідеально** — автори вимірюють це як AUC атаки **0,95 або вище** (0,5 — випадкове вгадування, 1,0 — безпомилковий результат), навіть коли середній показник по всьому датасету не кращий за випадковість.

Середній ризик може залишатися близьким до випадкового рівня, тоді як невеликий «хвіст» записів набагато сильніше викривається. Головна теза статті: приватність потрібно перевіряти на рівні окремого пацієнта, а не лише в сукупності.

Original Aurora diagram — The Clean Paper CC BY 4.0

Ризик розподілений нерівномірно — і припадає на недопредставлених. Пацієнти, найбільш уразливі до майже безпомилкової атаки, систематично належали до груп, **недопредставлених у даних**: етнічних меншин, рідкісних фенотипів хвороб, незвичайних характеристик зображень. У наборі електронних медичних записів темношкірі пацієнти траплялися серед найбільш уразливих записів **на 31% частіше**, ніж очікувалося; у наборі мамографії скани, позначені як підозрілі на злоякісність, були надпредставлені в цій зоні ризику на разючі **+1 179%**. (Ці два числа — приклади, а не вся картина: стаття описує той самий перекіс для кількох груп. І це *відносне* надпредставлення всередині малого «хвоста» крайнього ризику — тобто наскільки частіше такі пацієнти трапляються серед найбільш викритих записів, а не частка всіх темношкірих пацієнтів чи всіх пацієнтів із підозрілими мамограмами, які є викритими.) У моделі менше схожих прикладів, серед яких такі пацієнти могли б «загубитися», тому їхні записи сильніше вирізняються — а саме це й ловить membership inference. Провал приватності не випадковий; він концентрується на людях, які вже перебувають на маргінесі даних.

Більші моделі погіршують ситуацію. Кількість пацієнтів, уразливих до майже ідеальних атак, різко зростала разом із місткістю моделі. В одному дерматологічному датасеті частка піднялася від практично нуля в найменшій моделі до приблизно **одного з десяти** у найбільшій. Оскільки медичні моделі ШІ стають більшими й потужнішими — а саме в цьому напрямку рухається вся галузь, — цей конкретний

ризик, застерігають автори, посилюється, а не слабшає.

Підсумок Рюккерта різкий: «Це неприйнятний ризик. Медичні дані надзвичайно чутливі».

Чому «безпечна в середньому» — неправильна перевірка

Є спокуса прочитати низький середній ризик як довідку «все гаразд». Головний урок роботи полягає в тому, що усереднення тут не просто неточне — воно вимірює не те.

Гарантія приватності має сенс лише тоді, коли діє для людини з найбільшим ризиком, а не для людини посередині розподілу. Модель, у якій 999 із 1 000 пацієнтів неможливо ідентифікувати, а одного можна визначити майже ідеально, не є «приватною на 99,9%» у жодному сенсі, який має значення для цієї однієї людини. А якщо цією людиною передбачувано виявляється пацієнт із рідкісною хворобою або представник етнічної меншини, метрика не просто неповна — вона непомітно дискримінує. Вона вимірює безпеку більшості й називає це безпекою для всіх.

Саме такого зміщення вимагає стаття: від *наскільки приватна ця модель у середньому?* до *хто є найбільш викритим пацієнтом і до якої групи він належить?* Це різні запитання, і лише друге по-справжньому стосується гарантії приватності.

Чого ця робота *не* доводить — і не стверджує

- Вона **не** каже, що від медичного ШІ слід відмовитися. Рамка авторів — зменшення ризику, а не відступ: вимірювати й виправляти проблему, а не припиняти розробку.
- Вона **не** означає, що кожна медична модель щось витікає або що будь-яку конкретну розгорнуту модель вже атакували. Вона показує, що *ризик існує й розподілений нерівномірно*, а стандартні агреговані метрики його пропускають.
- Вона **не** означає, що ваші записи вже викриті. Атака потребує конкретних умов — доступу до моделі, запису-кандидата та власної інфраструктури зловмисника, — а не випадкової можливості будь-кого.
- Вона **не** показує витік *вмісту* картки пацієнта. Витікає належність — сам факт включення до навчальної вибірки. Це небезпечно з іншої причини, а не тому, що модель просто видає запис.
- Вона **не** зводить нерівність до одного гучного числа. Сильний і відтворюваний результат — це *патерн*: агреговані метрики занижують індивідуальний ризик, а найбільше його несуть недопредставлені пацієнти — в різних датасетах і сотнях моделей.

Наскільки переконливі докази?

Це емпіричний, методологічний результат, і він досить міцний. Картина — агреговані показники приватності систематично занижують ризик для окремого пацієнта, залишковий ризик концентрується на недопредставлених групах і зростає з місткістю моделі — повторилася на **семи датасетах різних типів** і великій кількості моделей для кожного. Саме така широта робить вимірювальне твердження переконливим, а не артефактом одного набору даних.

Є два чесні застереження. По-перше, це демонстрація *ризику*, виміряного атаками, які побудували самі автори. Вона характеризує, наскільки успішним *міг би* бути компетентний нападник за заданих припущень, а не те, як часто такі атаки реально відбуваються. По-друге, яскраві числа — майже ідеальний успіх атаки для деяких пацієнтів — навмисно описують людей із найгіршим результатом. У цьому й суть роботи, і читати ці числа треба як «хвіст розподілу набагато важчий, ніж підказує середнє», а не як «більшість пацієнтів викриті».

Правильна реакція — ані паніка, ані відмахування: це ретельне дослідження на багатьох наборах даних, яке показує, що стандартний спосіб сертифікувати приватність медичного ШІ сліпий до власних найгірших випадків — і що ці випадки припадають саме на пацієнтів, у яких і так найменше запасу безпеки.

Чому це важливо

Є дві причини, чому це більше, ніж технічна примітка.

По-перше, робота змінює те, що взагалі має право називатися «збереженням приватності». Якщо модель випускають із середнім показником приватності, цей показник може бути справді низьким і водночас приховувати підгрупу пацієнтів, яких membership inference майже безпомилково ідентифікує. Практична вимога статті конкретна: перед релізом оцінювати ризик приватності **для кожного пацієнта**, контролювати, хто має доступ до розгорнутих моделей, і використовувати методи на кшталт **диференційної приватності** — невеликий, ретельно відкалібрований шум під час навчання, який притупляє membership-атаки ціною реального й керованого зниження корисності моделі (компроміс між приватністю та корисністю тут явний, а не безкоштовний). «Ми перевірили середнє» більше не повинно вважатися достатньою перевіркою.

По-друге, приватність тут переплітається зі справедливістю. Ті самі групи, які недопредставлені в медичних даних — і тому медичний ШІ вже обслуговує їх гірше, — виявляються тими, чию приватність він захищає найслабше. Галузь, яка серйозно працює над першою нерівністю, не може вважати другу чужою проблемою. Йдеться про тих самих людей.

Заспокійлива історія про приватність медичного ШІ — *ми виміряли, середній ризик низький, отже все добре* — це саме та історія, яку ця стаття розбирає. Не для того,

щоб відлякати від технології, а щоб перемістити стандарт туди, де йому місце: це обіцянка, про виконання якої можна говорити лише після перевірки для людини, яку найімовірніше може бути скривджено.

Короткий підсумок

Атаки на визначення належності до навчальної вибірки намагаються з'ясувати, чи входив запис конкретної людини до навчальних даних моделі ШІ. Оскільки сама належність може розкрити чутливу інформацію — наприклад, що людина мала певну хворобу, — це справжній витік приватності навіть тоді, коли сам медичний запис ніколи не з'являється назовні. Раніше такі атаки оцінювали за тим, як часто вони успішні *в середньому* по датасету. У цій роботі ризик вимірювали *для кожного пацієнта* на семи медичних наборах даних (зображення, ЕКГ, електронні записи) та багатьох моделях для кожного. Виявилося, що агреговані метрики сильно занижують ризик: деяких людей можна ідентифікувати майже безпомилково (AUC атаки $\geq$ 0,95), навіть коли середній показник по датасету виглядає безпечним; найуразливішими систематично є представники недопредставлених груп (етнічні меншини, рідкісні хвороби, незвичайні зображення); а проблема посилюється зі збільшенням моделей. Автори не закликають відмовитися від медичного ШІ — вони пропонують вимірювати приватність на індивідуальному рівні, контролювати доступ до моделей і застосовувати диференційну приватність. Висновок: «приватна в середньому» — не гарантія приватності, а провал цієї гарантії найчастіше припадає на тих, хто вже захищений найменше.

Тверезa оцінка

Що показує стаття: На семи медичних наборах даних і багатьох моделях для кожного ризик membership inference для окремих пацієнтів набагато вищий, ніж підказують агреговані метрики — аж до майже ідеального успіху атаки (AUC $\geq$ 0,95). Цей залишковий ризик непропорційно припадає на недопредставлені групи й зростає з місткістю моделі.

Що правдоподібно, але не доведено: Що реальні зловмисники вже використовують це проти розгорнутих клінічних моделей. Дослідження демонструє *можливість і розподіл ризику* за заданих припущень, а не частоту атак у реальному світі.

Чого вона не показує: Що від медичного ШІ слід відмовитися; що кожна модель має витік або що будь-яку конкретну розгорнуту модель уже зламали; що витікає *вміст* картки пацієнта, а не факт належності; одного універсального числа нерівності — надійним результатом є патерн, а не одна цифра.

Головні обмеження: Робота вимірює ризик найгіршого випадку за допомогою атак самих авторів, а не спостережувані інциденти; драматичні числа навмисно описують

найбільш викритих людей; точні розбіжності між групами показані як послідовна картина в різних датасетах, а не зведені до одного числа.

Якої впевненості заслуговує висновок для загального читача? Високої, що агреговані метрики приватності занижують індивідуальний ризик, що залишковий ризик концентрується на недопредставлених пацієнтах і що він посилюється зі збільшенням моделей — це показано на багатьох датасетах і моделях. Помірної щодо реальної частоти таких атак, якої це дослідження не вимірює. Безпечне прочитання: не «медичний ШІ витікає ваші дані» і не «проблему приватності розв'язано», а «стандартна перевірка приватності сліпа до власних найгірших випадків, і ці випадки припадають на найуразливіших пацієнтів — отже перевірку треба змінити».

Джерела

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