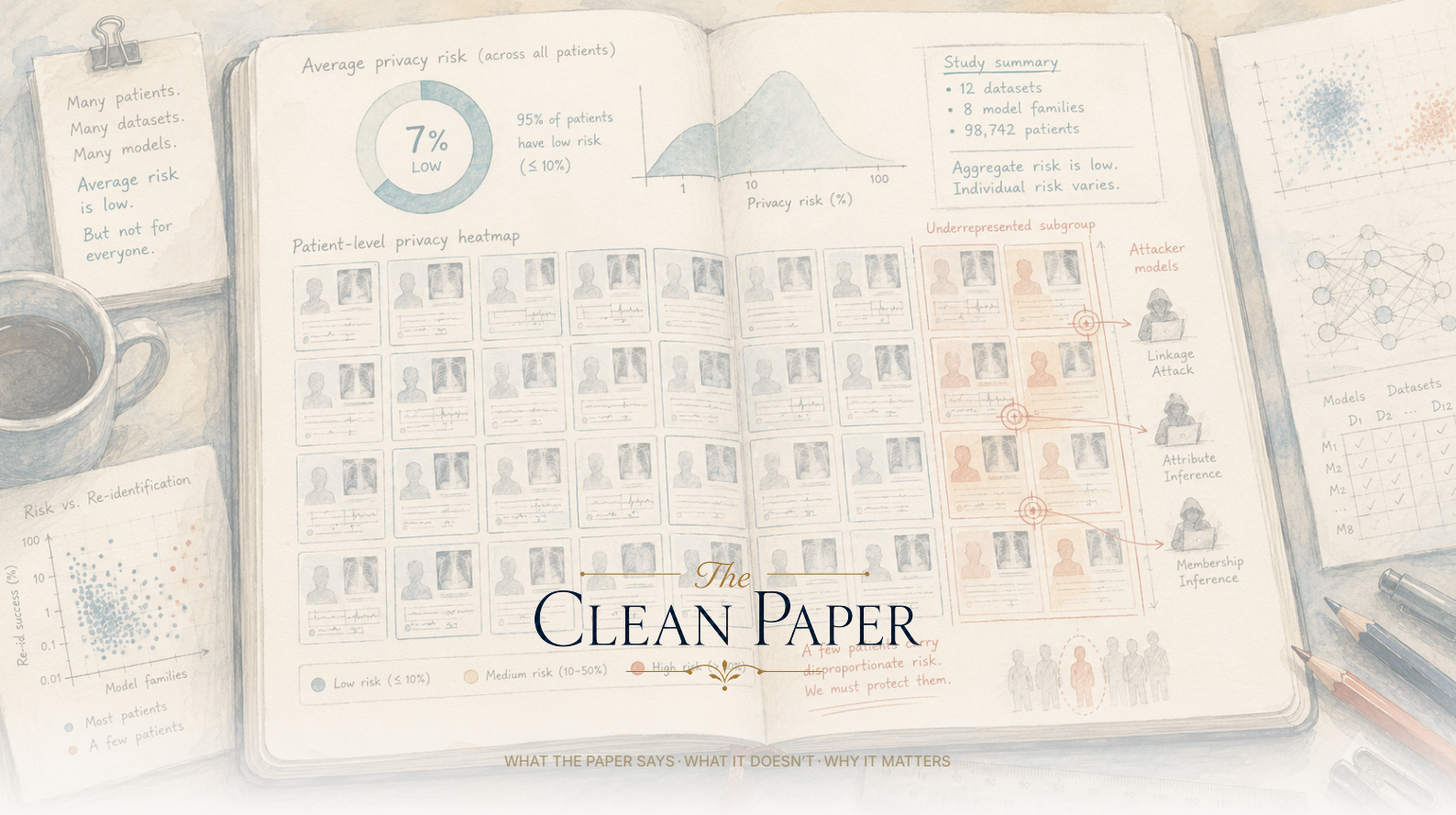


Медыцынскі ШІ можа быць прыватным у сярэднім і адначасова раскрываць асобных пацыентаў — асабліва недапрадстаўленых

Калі медыцынскую мадэль ШІ называюць «той, што захоўвае прыватнасць», гэта звычайна заснавана на адным сярэднім паказчыку. Гэтае даследаванне паказвае, што такая праверка няправільная. Вывучаючы атакі на вызначэнне прыналежнасці да навучальнай выбаркі — яны дазваляюць высветліць, ці быў запіс канкрэтнага чалавека ў навучальных даных мадэлі і тым самым могуць раскрыць, што ў чалавека было пэўнае захворванне, — аўтары вымяралі рызыку не ў сукупнасці, а для кожнага пацыента асобна, на сямі медыцынскіх наборах даных (выявы, ЭКГ, электронныя запісы) і мноства мадэляў. Карціна такая: мадэлі, бяспечныя паводле сярэдняга паказчыка, часам дазваляюць амаль беспамылкова ідэнтыфікаваць канкрэтных людзей (AUC атакі ≥ 0,95); сярод раскрываемых сістэматычна пераважаюць недапрадстаўленыя групы — этнічныя меншасці, рэдкія хваробы, незвычайныя выявы; а з павелічэннем мадэляў праблема ўзмацняецца. Аўтары не заклікаюць адмовіцца ад медыцынскага ШІ — яны прапануюць вымяраць прыватнасць для кожнага пацыента, кантраляваць доступ да мадэляў і ўжываць дыферэнцыяльную прыватнасць. Нязручная сутнасць: «прыватная ў сярэднім» — не гарантыя прыватнасці, і найчасцей яна падводзіць тых пацыентаў, якія і без таго абаронены найгорш.

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

Гарантыя прыватнасці — гэта абяцанне чалавеку, а не сярэдняму значэнню

Калі пра медыцынскую мадэль ШІ кажуць, што яна «захоўвае прыватнасць», звычайна гэта сцвярджэнне абапіраецца на адну лічбу: сярод усіх пацыентаў, на даных якіх навучалі мадэль, сярэдняя імавернасць вызначыць, ці ўваходзілі даныя канкрэтнага чалавека ў навучальны набор, невялікая. Гучыць супакойліва. Ціхая і нязручная выснова гэтай працы ў тым, што гэта проста не тая лічба, на якую трэба глядзець.

Прыватнасць — не сярэдняе значэнне. Гэта абяцанне кожнаму асобнаму чалавеку: удзел яго даных у навучальным наборы не павінен пазней яго раскрыць. А сярэдні паказчык можа выконваць гэтае абяцанне амаль для ўсіх і адначасова цалкам парушаць яго для некалькіх людзей. Даследчыкі вымералі рызыку для прыватнасці кожнага пацыента асобна, з дэталізацыяй, якой раней не ўжывалі, і ўбачылі менавіта гэта: мадэлі, што ў сукупнасці выглядаюць бяспечнымі, могуць амаль беспамылкова выдаваць сам факт удзелу канкрэтных людзей у навучальных даных — і непрапарцыйна часта гэта менавіта тыя людзі, якія і без таго абаронены найгорш.

Што зрабілі аўтары

Каманда пад кіраўніцтвам Морыца Кноле разам з Даніэлем Рукертам (абодва — Technical University of Munich) і Георгам Кайсісам (Hasso Plattner Institute), а таксама калегамі з Imperial College London узяла добра вядомую пагрозу прыватнасці — **атаку на вызначэнне прыналежнасці да навучальнай выбаркі (membership inference attack)** — і змяніла пытанне. Замест «як часта гэтая атака ў сярэднім паспяховая па ўсім наборы даных?» яны спыталі: «наколькі ўразлівы *гэты канкрэтны пацыент*?»

Такі аналіз на ўзроўні асобнага пацыента яны правялі на **сямі ўстояных медыцынскіх наборах даных**, якія ахопліваюць вельмі розныя тыпы інфармацыі — медыцынскія выявы, электракардыяграмы і электронныя медыцынскія запісы, — і для кожнага даследавалі па 200 мадэляў. Для кожнага пацыента ў кожнай мадэлі яны ацэньвалі, наколькі ўпэўнена зламыснік мог бы вызначыць, ці была карта гэтага чалавека часткай навучальных даных, а затым разбівалі вынікі па групах: статусе захворвання, самаідэнтыфікаванай расавай прыналежнасці, страхаванні, поле і пратаколе візуалізацыі.

> **Што насамрэч уяўляе сабой атака на вызначэнне прыналежнасці да навучальнай выбаркі**
>
> Уцечка тут больш тонкая, чым «мадэль выдае вашу медыцынскую карту». Гаворка пра *прыналежнасць* — сам факт таго, што вашы даныя былі ў навучальным наборы.
> Пачнём з таго, што такая мадэль робіць у звычайным рэжыме. Вы падаяце ёй выяву — скажам, рэнтген грудной клеткі — і атрымліваеце імавернасці: 78%

імавернасці пнеўманіі, а таксама ацэнкі для кардыямегаліі, ацёку, кансалідацыі. Менавіта гэты звычайны дыягнастычны адказ і выкарыстоўвае атака. Нічога экзатычнага.

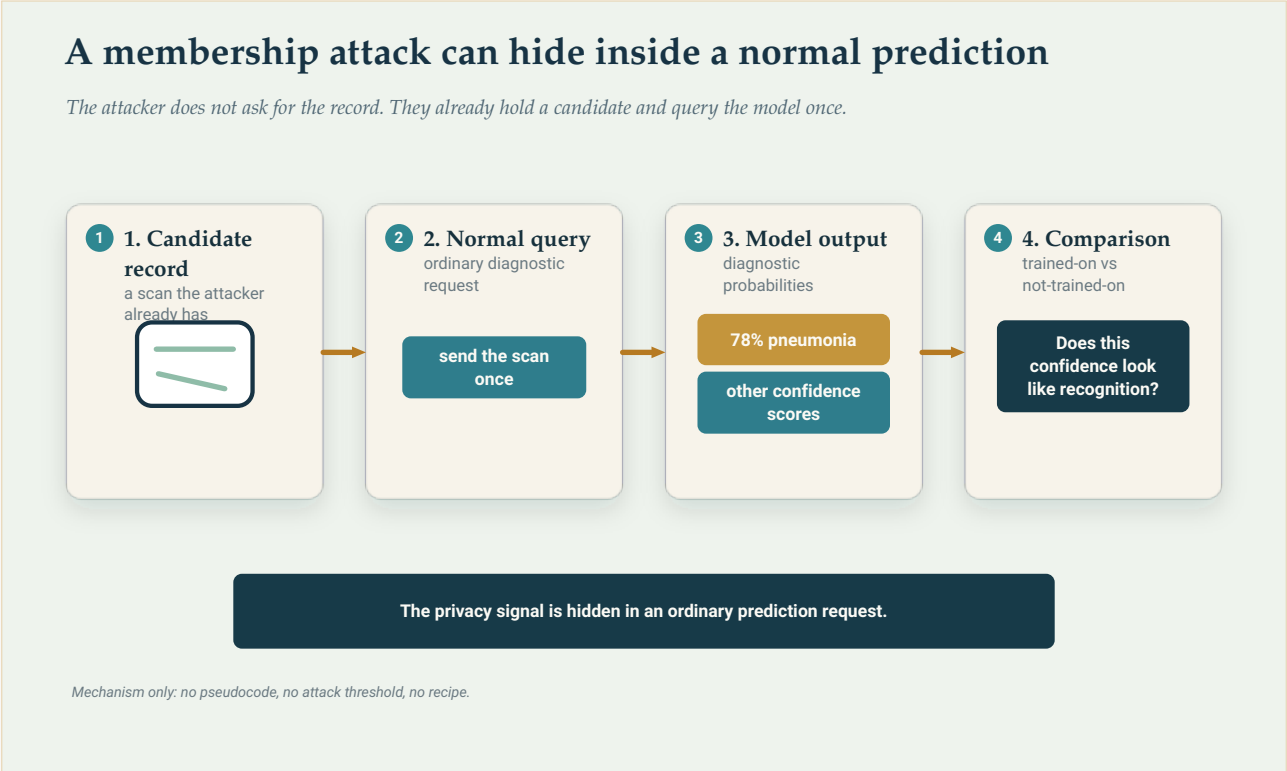
Слабае месца вось у чым: мадэль звычайна крыху **больш упэўненая** менавіта на тых прыкладах, на якіх яе навучалі, чым на тых, якіх яна ніколі не бачыла. Уявіце студэнта, які ўпотай загадзя бачыў экзаменацыйныя пытанні: на ўжо адпрацаваныя пытанні адказы прыходзяць крыху хутчэй і крыху больш упэўнена. Вызначыць па гэтай прыкмеце, ці быў канкрэтны запіс сярод прыкладаў, якія мадэль «вывучала», — гэта і ёсць вызначэнне прыналежнасці да навучальнай выбаркі.

Такім чынам, атака выглядае так. Нехта хоча даведацца, ці выкарыстоўваўся скан канкрэтнага чалавека для навучання мадэлі. Ён **не** просіць мадэль вярнуць запіс — у яго ўжо ёсць скан-кандыдат, уласны скан гэтага чалавека або вельмі блізкая копія. Ён адзін раз падае яго як звычайны дыягнастычны запыт і глядзіць, наколькі ўпэўнена адказвае мадэль. Затым параўноўвае гэтую ўпэўненасць з паводзінамі мадэлі, якая *навучалася* на такім скане, і мадэлі, якая *не навучалася* на ім. Такое параўнанне можна нядорага пабудаваць, навучыўшы ўласную дапаможную «рэферэнтную мадэль», каб яна паказала характэрную залішнюю ўпэўненасць; для гэтага дастаткова звычайнага камп'ютара без спецыяльнага абсталявання. Калі сапраўдная мадэль незвычайна ўпэўненая менавіта на гэтым скане — быццам пазнае яго, — гэта моцны сігнал, што скан уваходзіў у навучальныя даныя.

Самае непрыемнае, што сам запыт зусім не выглядае як атака: гэта той самы дыягнастычны запыт, які мог бы зрабіць клініцыст, а сігнал пра прыватнасць схаваны ўнутры звычайнага прагнозу. Таму рызыка цалкам канкрэтная — хоць пакуль яе дэманстравалі пры выразна вызначаных лабараторных дапушчэннях, а не назіралі ў рэальных атаках.

Навошта ўвогуле важная прыналежнасць, калі сам запіс ніколі не ўцякае? Таму што *прыналежнасць таксама з'яўляецца фактам*. Калі мадэль навучалі на пацыентах, якія атрымлівалі пэўную імунатэрапію раку, пацвярджэнне таго, што ваш запіс уваходзіў у гэты набор, можа раскрыць, што ў вас, верагодна, быў гэты рак — менавіта той тып інфармацыі, які страхоўшчык або працадаўца не павінен мець магчымасці вывесці. Змесціва запісу застаецца закрытым; вонкі выходзіць факт прыналежнасці да групы.

Аўтары проста апісваюць перадумовы атакі — доступ да звычайных прагнозаў мадэлі, наяўнасць запісу-кандыдата і ўласнай рэферэнтнай мадэлі зламысніка — не як інструкцыю, а каб пагрозу можна было аналізаваць канкрэтна, а не адмахвацца ад яе.



Для атакі не патрэбны спецыяльны API прыватнасці. Звычайны дыягнастычны запыт можа несці сігнал пра прыналежнасць да навучальнай выбаркі, калі мадэль незвычайна ўпэўненая на запісе, які ўжо бачыла.

Што яны выявілі

Тры вынікі, кожны вастрэйшы за папярэдні.

Сярэднія значэнні хаваюць найбольш уразлівых. Калі вымяраць у сукупнасці — як звычайна і робяць, — многія мадэлі выглядаюць дастаткова прыватнымі: для большасці пацыентаў атака працуе не лепш за выпадковае адгадванне. Але на ўзроўні асобнага чалавека карціна іншая. Як Кноле сказаў у паведамленні пра даследаванне, папярэднія ацэнкі «заўсёды вымяралі толькі сярэднюю рызыку для ўсіх пацыентаў. Мы ўпершыню даследавалі рызыку на ўзроўні асобных пацыентаў — і карціна аказалася зусім іншай». Для некаторых людзей атака працуе **амаль ідэальна** — аўтары вымяраюць гэта як AUC атакі **0,95 або вышэй** (0,5 — выпадковае адгадванне, 1,0 — беспамылковы вынік), нават калі сярэдні паказчык па ўсім наборы даных не лепшы за выпадковасць.

Сярэдняя рызыка можа заставацца блізкай да выпадковага ўзроўню, тады як невялікі «хвост» запісаў раскрываецца значна мацней. Галоўны тэзіс артыкула: прыватнасць трэба правяраць на ўзроўні асобнага пацыента, а не толькі ў сукупнасці.

Рызыка размеркавана нераўнамерна — і прыпадае на недапрадстаўленых. Пацыенты, найбольш уразлівыя для амаль беспамылковай атакі, сістэматычна належалі да груп, **недапрадстаўленых у даных**: этнічных меншасцей, рэдкіх фенатыпаў захворванняў, незвычайных характарыстык выяў. У наборы электронных медыцынскіх запісаў чарнаскурыя пацыенты сустракаліся сярод найбольш уразлівых запісаў **на 31% часцей**, чым чакалася; у наборы мамаграфіі сканы, пазначаныя як падазроныя на злаякаснасць, былі звышпрадстаўленыя ў гэтай зоне рызыкі на ўражальныя **+1 179%**. (Гэтыя дзве лічбы — прыклады, а не ўся карціна: артыкул апісвае той самы перакос для некалькіх груп. І гэта *адноснае* звышпрадстаўленне ўнутры малога «хваста» крайняй рызыкі — гэта значыць, наколькі часцей такія пацыенты сустракаюцца сярод найбольш раскрываемых запісаў, а не доля ўсіх чарнаскурых пацыентаў або ўсіх пацыентаў з падазронымі мамаграмамі, якія раскрываюцца.) У мадэлі менш падобных прыкладаў, сярод якіх такія пацыенты маглі б «згубіцца», таму іх запісы мацней вылучаюцца — а менавіта гэта і ўлоўлівае атака на прыналежнасць да навучальнай выбаркі. Правал прыватнасці не выпадковы; ён канцэнтруецца на людзях, якія ўжо знаходзяцца на перыферыі даных.

Буйнейшыя мадэлі пагаршаюць сітуацыю. Колькасць пацыентаў, уразлівых для амаль ідэальных атак, рэзка расла разам з ёмістасцю мадэлі. У адным дэрматалагічным наборы даных доля паднялася ад практычна нуля ў найменшай мадэлі прыкладна да **аднаго з дзесяці** ў найбольшай. Паколькі медыцынскія мадэлі ШІ становяцца

буйнейшымі і магутнейшымі — а менавіта ў гэтым кірунку рухаецца ўся галіна, — гэтая канкрэтная рызыка, папярэджваюць аўтары, узмацняецца, а не слабее.

Выснова Рукерта рэзкая: «Гэта непрымальная рызыка. Медыцынскія даныя надзвычай адчувальныя».

Чаму «бяспечная ў сярэднім» — няправільная праверка

Ёсць спакуса прачытаць нізкую сярэднюю рызыку як даведку «ўсё ў парадку». Галоўны ўрок працы ў тым, што асерадненне тут не проста недакладнае — яно вымярае не тое.

Гарантыя прыватнасці мае сэнс толькі тады, калі дзейнічае для чалавека з найбольшай рызыкай, а не для чалавека ў сярэдзіне размеркавання. Мадэль, у якой 999 з 1 000 пацыентаў немагчыма ідэнтыфікаваць, а аднаго можна вызначыць амаль ідэальна, не з'яўляецца «прыватнай на 99,9%» ні ў якім сэнсе, важным для гэтага аднаго чалавека. А калі гэтым чалавекам прадказальна аказваецца пацыент з рэдкай хваробай або прадстаўнік этнічнай меншасці, метрыка не проста няпоўная — яна незаўважна дыскрымінуе. Яна вымярае бяспеку большасці і называе гэта бяспекай для ўсіх.

Менавіта такога зруху патрабуе артыкул: ад *наколькі прыватная гэтая мадэль у сярэднім?* да *хто з'яўляецца найбольш раскрываемым пацыентам і да якой групы ён належыць?* Гэта розныя пытанні, і толькі другое сапраўды датычыцца гарантыі прыватнасці.

Чаго гэтая праца *не* даказвае — і не сцвярджае

- Яна **не** кажа, што ад медыцынскага ШІ трэба адмовіцца. Падыход аўтараў — зніжэнне рызыкі, а не адступленне: вымяраць і выпраўляць праблему, а не спыняць распрацоўку.
- Яна **не** азначае, што кожная медыцынская мадэль нешта ўцеча або што нейкую канкрэтную разгорнутую мадэль ужо атакавалі. Яна паказвае, што *рызыка існуе і размеркавана нераўнамерна*, а стандартныя агрэгаваныя метрыкі яе прапускаюць.
- Яна **не** азначае, што вашы запісы ўжо раскрытыя. Атака патрабуе канкрэтных умоў — доступу да мадэлі, запісу-кандыдата і ўласнай інфраструктуры зламысніка, — а не выпадковай магчымасці любога чалавека.
- Яна **не** паказвае ўцечку *змесціва* карты пацыента. Уцякае прыналежнасць — сам факт уключэння ў навучальную выбарку. Гэта небяспечна па іншай прычыне, а не таму, што мадэль проста выдае запіс.
- Яна **не** зводзіць няроўнасць да адной гучнай лічбы. Моцны і ўзнаўляльны вынік — гэта *заканамернасць*: агрэгаваныя метрыкі заніжаюць індывідуальную рызыку, а найбольшую рызыку нясуць недапрадстаўленыя пацыенты — у розных наборах даных і сотнях мадэляў.

Наколькі пераканаўчыя сведчанні?

Гэта эмпірычны метадалагічны вынік, і ён дастаткова трывалы. Карціна — агрэгаваныя паказчыкі прыватнасці сістэматычна заніжаюць рызыку для асобнага пацыента, рэшткавая рызыка канцэнтруецца на недапрадстаўленых групах і расце з ёмістасцю мадэлі — паўтарылася на **сямі наборах даных розных тыпаў** і вялікай колькасці мадэляў для кожнага. Менавіта такая шырыня робіць сцвярджэнне пра праблему вымярэння пераканаўчым, а не артэфактам аднаго набору даных.

Ёсць дзве сумленныя агаворкі. Па-першае, гэта дэманстрацыя *рызыкі*, вымеранай атакамі, якія пабудавалі самі аўтары. Яна характарызуе, наколькі паспяховым *мог бы* быць кампетэнтны зламыснік пры зададзеных дапушчэннях, а не тое, як часта такія атакі сапраўды адбываюцца. Па-другое, яркія лічбы — амаль ідэальны поспех атакі для некаторых пацыентаў — наўмысна апісваюць людзей з найгоршым вынікам. У гэтым і ёсць сэнс працы, і чытаць гэтыя лічбы трэба як «хвост размеркавання значна цяжэйшы, чым падказвае сярэдняе», а не як «большасць пацыентаў раскрытыя».

Правільная рэакцыя — ні паніка, ні адмахванне: гэта дбайнае даследаванне на мностве набораў даных, якое паказвае, што стандартны спосаб сертыфікаваць прыватнасць медыцынскага ШІ сляпы да ўласных найгоршых выпадкаў — і што гэтыя выпадкі прыпадаюць менавіта на пацыентаў, у якіх і так найменшы запас бяспекі.

Чаму гэта важна

Ёсць дзве прычыны, чаму гэта больш, чым тэхнічная заўвага.

Па-першае, праца змяняе тое, што ўвогуле мае права называцца «захаваннем прыватнасці». Калі мадэль выпускаюць з сярэднім паказчыкам прыватнасці, гэты паказчык можа быць сапраўды нізкім і пры гэтым хаваць падгрупу пацыентаў, для якіх атака на прыналежнасць да навучальнай выбаркі амаль беспамылкова спрацоўвае. Практычнае патрабаванне артыкула канкрэтнае: перад выпускам ацэньваць рызыку для прыватнасці **кожнага пацыента**, кантраляваць, хто мае доступ да разгорнутых мадэляў, і выкарыстоўваць метады кшталту **дыферэнцыяльнай прыватнасці** — невялікі, старанна адкалібраваны шум падчас навучання, які аслабляе атакі на прыналежнасць да навучальнай выбаркі коштам рэальнага і кіраванага зніжэння карыснасці мадэлі (кампраміс паміж прыватнасцю і карыснасцю тут відавочны, а не бясплатны). «Мы праверылі сярэдняе» больш не павінна лічыцца дастатковай праверкай.

Па-другое, прыватнасць тут пераплятаецца са справядлівасцю. Тыя ж групы, якія недапрадстаўлены ў медыцынскіх даных — і таму медыцынскі ШІ ўжо абслугоўвае іх горш, — аказваюцца тымі, чыю прыватнасць ён абараняе найслабей. Галіна, якая сур’ёзна працуе над першай няроўнасцю, не можа лічыць другую чужой праблемай. Гаворка пра тых самых людзей.

Супакойлівая гісторыя пра прыватнасць медыцынскага ШІ — *мы вымералі, сярэдняя рызыка нізкая, значыць усё добра* — менавіта тая гісторыя, якую гэты артыкул разбірае. Не для таго, каб адпужваць ад тэхналогіі, а каб перамясціць стандарт туды, дзе яму месца: гэта абяцанне, пра выкананне якога можна гаварыць толькі пасля праверкі для чалавека, якому з найбольшай імавернасцю можа быць нанесена шкода.

Коратка

Атакі на вызначэнне прыналежнасці да навучальнай выбаркі спрабуюць высветліць, ці ўваходзіў запіс канкрэтнага чалавека ў навучальныя даныя мадэлі ШІ. Паколькі сама прыналежнасць можа раскрыць адчувальную інфармацыю — напрыклад, што ў чалавека было пэўнае захворванне, — гэта сапраўдная ўцечка прыватнасці нават тады, калі сам медыцынскі запіс ніколі не выходзіць вонкі. Раней такія атакі ацэньвалі па тым, як часта яны паспяховыя *ў сярэднім* па наборы даных. У гэтай працы рызыку вымяралі *для кожнага пацыента* на сямі медыцынскіх наборах даных (выявы, ЭКГ, электронныя запісы) і мноства мадэляў для кожнага. Аказалася, што агрэгаваныя метрыкі моцна заніжаюць рызыку: некаторых людзей можна ідэнтыфікаваць амаль беспамылкова (AUC атакі $\geq$ 0,95), нават калі сярэдні паказчык па наборы выглядае бяспечным; найбольш уразлівыя сістэматычна прадстаўнікі недапрадстаўленых груп (этнічныя меншасці, рэдкія хваробы, незвычайныя выявы); а праблема ўзмацняецца па меры павелічэння мадэляў. Аўтары не заклікаюць адмовіцца ад медыцынскага ШІ — яны прапануюць вымяраць прыватнасць на індывідуальным узроўні, кантраляваць доступ да мадэляў і ўжываць дыферэнцыяльную прыватнасць. Выснова: «прыватная ў сярэднім» — не гарантыя прыватнасці, а правал гэтай гарантыі найчасцей прыпадае на тых, хто і без таго абаронены найгорш.

Праверка без прыкрас

Што паказвае артыкул: На сямі медыцынскіх наборах даных і мноства мадэляў для кожнага рызыка атакі на прыналежнасць да навучальнай выбаркі для асобных пацыентаў значна вышэйшая, чым падказваюць агрэгаваныя метрыкі, — аж да амаль ідэальнага поспеху атакі (AUC $\geq$ 0,95). Гэтая рэшткавая рызыка непрапарцыйна прыпадае на недапрадстаўленыя групы і расце з ёмістасцю мадэлі.

Што праўдападобна, але не даказана: Што рэальныя зламыснікі ўжо выкарыстоўваюць гэта супраць разгорнутых клінічных мадэляў. Даследаванне дэманструе *магчымасць і размеркаванне рызыкі* пры зададзеных дапушчэннях, а не частату атак у рэальным свеце.

Чаго яно не паказвае: Што ад медыцынскага ШІ трэба адмовіцца; што кожная мадэль мае ўцечку або нейкую канкрэтную разгорнутую мадэль ужо ўзламалі; што ўцякае

змесціва карты пацыента, а не факт прыналежнасці; аднаго ўніверсальнага ліку няроўнасці — надзейны вынік тут заканамернасць, а не адна лічба.

Галоўныя абмежаванні: Праца вымярае рызыку найгоршага выпадку з дапамогай атак саміх аўтараў, а не назіраныя інцыдэнты; драматычныя лічбы наўмысна апісваюць найбольш раскрываемых людзей; дакладныя адрозненні паміж групамі паказаны як паслядоўная карціна ў розных наборах даных, а не зведзены да адной лічбы.

Якой упэўненасці заслугоўвае выснова для шырокага чытача? Высокай у тым, што агрэгаваныя метрыкі прыватнасці заніжаюць індывідуальную рызыку, што рэшткавая рызыка канцэнтруецца на недапрадстаўленых пацыентах і ўзмацняецца з павелічэннем мадэляў — гэта паказана на многіх наборах даных і мадэлях. Умеранай адносна рэальнай частаты такіх атак, якую гэтае даследаванне не вымярае. Бяспечнае прачытанне: не «медыцынскі ШІ злівае вашы даныя» і не «праблема прыватнасці вырашана», а «стандартная праверка прыватнасці сляпая да ўласных найгоршых выпадкаў, і гэтыя выпадкі прыпадаюць на найбольш уразлівых пацыентаў — значыць, праверку трэба змяніць».

Крыніцы

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