The Clean Paper

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

AI ya kitabibu inaweza kuwa private kwa wastani lakini bado ikafichua wagonjwa fulani — hasa waliowakilishwa kidogo

Modeli ya AI ya kitabibu inayoitwa 'privacy-preserving' mara nyingi hutegemea namba moja ya wastani. Utafiti huu unasema hiyo ndiyo test isiyofaa. Kwa kuchunguza membership inference attacks — ambazo hufichua kama rekodi ya mtu fulani ilikuwa katika training data na hivyo zinaweza kusaliti kwamba alikuwa na ugonjwa fulani — waandishi walipima hatari kwa kila mgonjwa badala ya kwa jumla, katika datasets saba za kitabibu na modeli nyingi. Pattern: modeli zinazonekana salama kwa wastani bado zinaweza kuruhusu baadhi ya watu kutambuliwa karibu kikamilifu (attack AUC $\geq$ 0.95); walio wazi zaidi ni kwa utaratibu makundi yaliyowakilishwa kidogo; na tatizo huzidi kadiri modeli zinavyokuwa kubwa. Waandishi hawasemi AI iachwe — wanasema faragha ipimwe kwa kila mgonjwa, access kwa modeli idhibitiwe na differential privacy itumike.

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

Ahadi ya faragha ni ahadi kwa mtu, si kwa wastani

Modeli ya AI ya kitabibu inapoitwa “inalinda faragha,” dai hilo mara nyingi hutegemea namba moja: kwa wastani katika wagonjwa wote ambao data zao zilitumika kuifundisha, uwezekano wa kubaini kama rekodi ya mtu fulani ilikuwa ndani ya seti ya mafunzo ni mdogo. Hilo linasikika kutia moyo. Hoja tulivu lakini isiyofurahisha ya makala hii ni kwamba hiyo ndiyo namba isiyofaa kuangalia.

Faragha si wastani. Ni ahadi kwa kila mtu — kwamba kuwapo kwake katika seti ya mafunzo hakutarudi kumfichua. Wastani unaweza kutimiza ahadi hiyo kwa karibu kila mtu huku ukiivunja kabisa kwa wachache. Watafiti walipima hatari ya faragha mgonjwa mmoja baada ya mwingine, kwa resolution ambayo haikuwa imetumiwa hapo awali, na wakapata hilo hasa: modeli zinazoonekana salama kwa jumla zinaweza kufichua membership ya watu fulani karibu kikamilifu — na watu wanaofichuliwa kwa kiwango kisicho sawa ni wale ambao tayari wako katika makundi yenye ulinzi mdogo zaidi.

Waandishi walifanya nini

Timu — ikiongozwa na Moritz Knolle, pamoja na Daniel Rückert (wote Technical University of Munich) na Georg Kaissis (Hasso Plattner Institute), pamoja na wenzao Imperial College London — ilichukua tishio linalojulikana la faragha linaloitwa **membership inference attack** na kubadilisha swali. Badala ya kuuliza “kwa wastani, shambulio hili linafanikiwa mara ngapi katika hifadhidata?”, waliuliza “kwa *mgonjwa huyu hasa*, yuko wazi kiasi gani?”

Walifanya uchambuzi huo kwa kila mgonjwa katika **hifadhidata saba zinazotumika sana za kitabibu**, zenye aina tofauti sana za data—picha za kitabibu, electrocardiograms na electronic health records—na kwa kila hifadhidata wakachunguza modeli 200. Kwa kila mgonjwa katika kila modeli, walikadiria kwa uhakika kiasi gani mshambulizi angeweza kujua kama rekodi ya mtu huyo ilikuwa sehemu ya data ya mafunzo, kisha wakagawa matokeo kwa makundi kama hali ya ugonjwa, race iliyoripotiwa na mtu mwenyewe, bima, jinsia na utaratibu wa upigaji picha.

Membership inference attack ni nini hasa

Uvujaji hapa ni wa hila zaidi kuliko “modeli inatoa rekodi yako.” Kinachovuja ni *membership* — ukweli tu kwamba data yako ilikuwa ndani ya seti ya mafunzo.

Anza na kazi ya kawaida ya modeli kama hizi. Unaipa scan — kwa mfano X-ray ya kifua — nayo inarudisha probabilities: uwezekano wa 78% wa pneumonia, pamoja na makadirio ya cardiomegaly, oedema, consolidation. Jibu hilo la kawaida la utambuzi ndilo *pekee* ambalo shambulio linatumia. Hakuna API ya ajabu.

Udhaifu unaotumiwa ni huu: kwa kawaida modeli huwa **na uhakika kidogo zaidi** juu ya mifano halisi iliyotumiwa kuifundisha kuliko mifano ambayo haijawahi kuona. Fikiria mwanafunzi aliyeyaona maswali ya mtihani kwa siri kabla — kwenye maswali aliyojizoeza, majibu yanatoka haraka kidogo na kwa uhakika mwingi kuliko kawaida. Kutambua kutokana na ishara hiyo

kama rekodi fulani ilikuwa miongoni mwa mifano ambayo modeli ilisoma ndiyo maana ya “membership inference.”

Kwa hiyo shambulio hufanya kazi hivi. Mtu anataka kujua kama picha ya uchunguzi ya mtu fulani ilitumika kufundisha modeli. **Haombi** modeli itoe rekodi — tayari ana picha hiyo. Anaituma kama ombi la kawaida la utambuzi na kuangalia kiwango cha uhakika cha modeli. Kisha analinganisha kama uhakika huo unafanana zaidi na tabia ya modeli iliyowahi kufundishwa kwa picha hiyo au ya modeli ambayo haikuiona wakati wa mafunzo. Anaweza kujifunza tofauti hiyo kwa gharama ndogo kwa kufundisha modeli za rejea. Ikiwa modeli halisi ina uhakika usio wa kawaida juu ya picha hiyo — kana kwamba “inaikumbuka” — hilo linaongeza ushahidi kwamba picha ilikuwa katika data ya mafunzo.

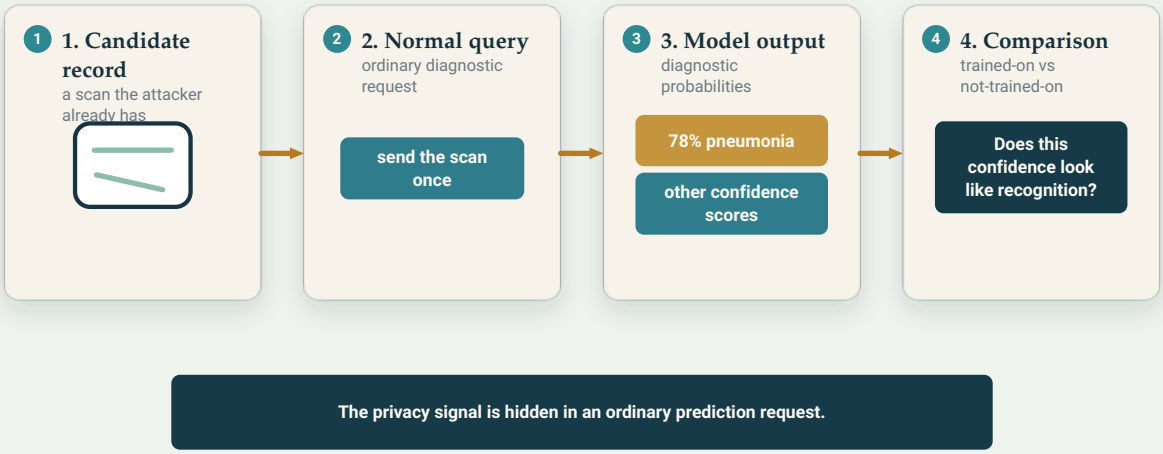
Sehemu inayotia wasiwasi ni kwamba ombi lenyewe halionekani kama shambulio: ni swali lilelile la utambuzi ambalo clinician angeuliza, na ishara ya faragha imejificha ndani ya prediction ya kawaida. Ndiyo maana hatari ni halisi — ingawa hadi sasa imeonyeshwa chini ya dhana za maabara zilizoelezwa, si kunaswa katika shambulio la dunia halisi.

Kwa nini membership ni muhimu kama rekodi yenyewe haitoki? Kwa sababu *membership yenyewe ni taarifa*. Kama modeli ilifundishwa kwa wagonjwa waliopata immunotherapy fulani ya saratani, kuthibitisha kwamba rekodi yako iko ndani yake kunaweza kufichua kwamba huenda ulikuwa na saratani hiyo — taarifa ambayo kampuni ya bima au mwajiri hapaswi kuweza kuhitimisha. Maudhui ya rekodi yanabaki yamefungwa; ukweli wa kuwa ndani ndiyo unaovuja.

Waandishi wanaweka wazi dhana hizi — access kwa predictions za kawaida za modeli, rekodi inayoshukiwa, na reference modeli ya mshambulizi — si kama mwongozo wa kushambulia, bali ili tishio lijadiliwe kwa msingi badala ya kupuuzwa kwa ishara ya mkono.

A membership attack can hide inside a normal prediction

The attacker does not ask for the record. They already hold a candidate and query the model once.



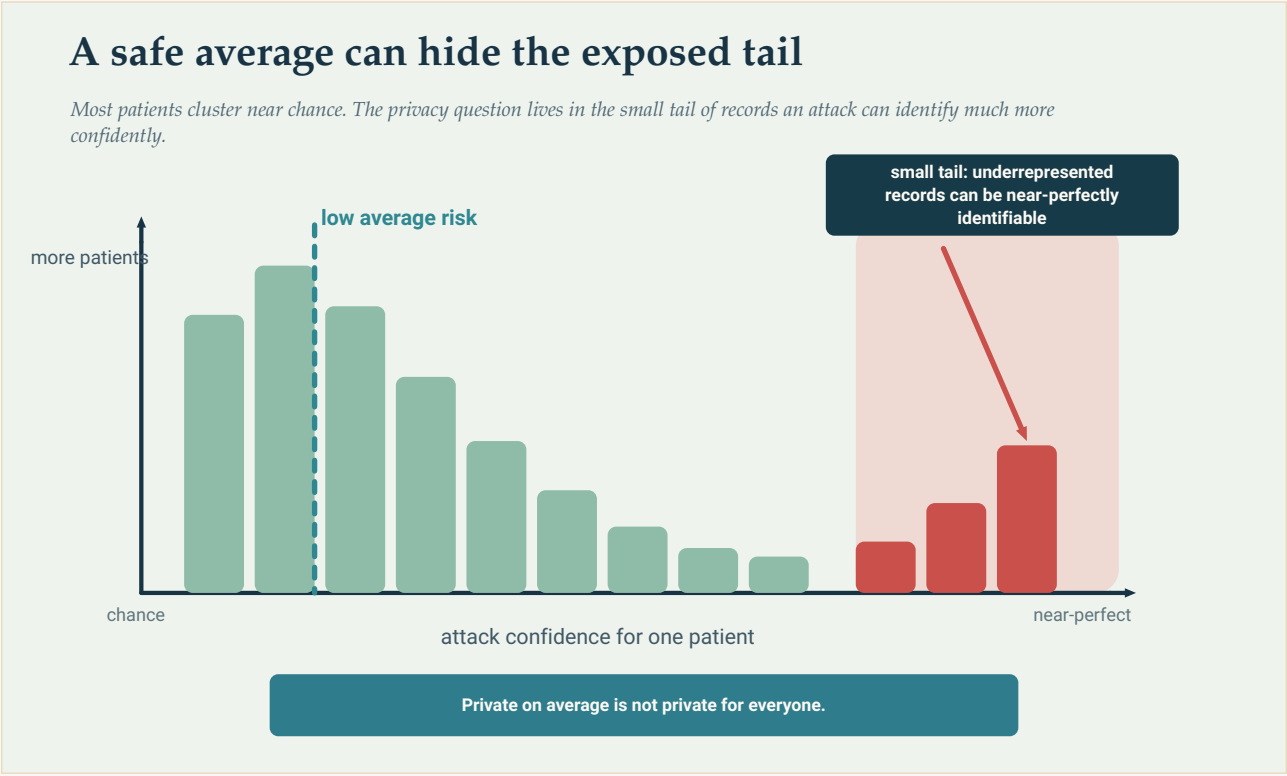
Mechanism only: no pseudocode, no attack threshold, no recipe.

Shambulio halihitaji API maalum ya faragha. Ombi la kawaida la utambuzi linaweza kuvujisha ishara ya membership kama modeli ina confidence isiyo ya kawaida juu ya rekodi ambayo imewahi kuona.

Walichokuta

Matokeo matatu, kila moja likiwa kali zaidi kuliko lililotangulia.

Wastani huficha walio wazi. Ukipimwa kwa jumla — njia ya kawaida — modeli nyingi zinaonekana kutia moyo: kwa wagonjwa wengi shambulio halifanyi vizuri kuliko kubahatisha. Mtazamo wa kila mgonjwa ulitoa picha tofauti. Kama Knolle alivyosema katika tangazo la utafiti, tathmini za awali "zimewahi kupima hatari ya wastani tu kwa wagonjwa wote. Sisi tulichunguza hatari katika kiwango cha mgonjwa mmoja mmoja kwa mara ya kwanza — na picha ni tofauti sana." Kwa baadhi ya watu, shambulio linafanikiwa **karibu kikamilifu** — waandishi wanapima hii kama attack AUC ya **0.95 au zaidi** (0.5 ni kama kurusha sarafu, 1.0 ni ukamilifu) — hata wakati wastani wa hifadhidata nzima unaonekana si bora kuliko kubahatisha.



Wastani unaweza kukaa karibu na kiwango cha bahati huku mkia mdogo wa rekodi ukiwa wazi zaidi kwa utambuzi. Hoja ya makala ni kwamba faragha lazima ikaguliwe katika kiwango cha mgonjwa, si kwa jumla pekee.

Uwazi huo haujasambazwa sawa — na unawaangukia zaidi waliowakilishwa kidogo. Wagonjwa walio katika hatari kubwa zaidi ya kutambuliwa walikuwa mara nyingi kutoka makundi **yasiyowakil-**

ishwa vya kutosha katika data: makabila madogo, aina nadra za magonjwa au sifa zisizo za kawaida katika picha za kitabibu. Katika hifadhidata ya rekodi za kielektroniki, wagonjwa Weusi walionekana **31% mara nyingi zaidi** kuliko ilivyotarajiwa miongoni mwa rekodi zilizo hatarini zaidi; katika hifadhidata ya mammography, picha zilizotiwa alama kuwa na shaka ya saratani zilikuwa zimewakilishwa kupita kiasi katika mkia huo wa hatari kwa **+1,179%**. Namba hizo ni mifano, si picha nzima, na zinaonyesha uwakilishi wa ziada *ndani ya kundi dogo lenye hatari kubwa*, si asilimia ya wagonjwa wote wa kundi hilo waliofichuliwa. Modeli ina mifano michache inayofanana nao, hivyo rekodi zao huwa rahisi kujitokeza. Kushindwa kwa faragha kwa hiyo si kwa nasibu; kunaweza kujikusanya kwa watu ambao tayari wako pembezoni mwa data.

Modeli kubwa zaidi zinafanya hali iwe mbaya. Idadi ya wagonjwa walio wazi kwa mashambulio karibu kamili ilipanda kwa kasi kadiri uwezo wa modeli ulivyoongezeka — katika hifadhidata moja ya dermatology, sehemu hiyo ilipanda kutoka karibu sifuri katika modeli ndogo zaidi hadi takribani **mmoja kati ya kumi** katika modeli kubwa zaidi. Kadiri modeli za AI ya kitabibu zinavyokuwa kubwa na zenye uwezo zaidi — mwelekeo wa uwanja mzima — waandishi wanaonya kwamba hatari hii maalum inakuwa kubwa zaidi, si ndogo.

Muhtasari wa Rückert ni mkali: “Hii si hatari inayoweza kuvumiliwa. Data ya afya ni nyeti sana.”

Kwa nini “salama kwa wastani” ni kipimo kibaya

Ni rahisi kusoma hatari ndogo ya wastani kama cheti safi cha afya. Somo kuu la makala ni kwamba averaging hapa si kukosa usahihi tu — ni kupima kitu kisicho sahihi.

Ahadi ya faragha ina maana tu ikiwa inashikilia kwa mtu aliye katika hatari kubwa zaidi, si mtu wa katikati. Modeli ambayo wagonjwa 999 kati ya 1,000 hawawezi kutambuliwa lakini mmoja anaweza kutambuliwa karibu kikamilifu si “99.9% private” kwa maana yoyote inayomsaidia mtu huyo mmoja — na kama mtu huyo ana uwezekano mkubwa wa kuwa mgonjwa mwenye ugonjwa nadra au wa kundi la kikabila lililo wachache, metric hiyo si pungufu tu, bali ina athari ya ubaguzi kimya kimya. Inaripoti usalama wa walio wengi na kuuita usalama wa wote.

Hilo ndilo badiliko makala inalazimisha: kutoka *modeli hii ni private kiasi gani kwa wastani?* hadi *ni mgonjwa gani aliye wazi zaidi, na anatoka kundi gani?* Hayo ni maswali tofauti, na la pili ndilo swali halisi la faragha.

Kile ambacho hii *haithibitishi* — wala kudai

- **Haisemi** AI ya kitabibu iachwe. Mtazamo wa waandishi ni kupunguza hatari, si kurudi nyuma: pima na rekebisha hatari, usiache kujenga.
- **Haimaanishi** kila modeli ya kitabibu inavuja, au kwamba modeli fulani iliyowekwa kazini tayari imeshambuliwa. Inaonyesha *hatari ipo na haijasambazwa sawa*, na vipimo za kawaida za jumla haziioni vizuri.

- **Haimaanishi** rekodi zako tayari ziko wazi. Shambulio linahitaji hali maalum — access kwa modeli, rekodi inayoshukiwa na miundombinu ya mshambulizi — si uwezo wa kawaida wa mtu yeyote.
- **Haionyeshi** maudhui ya rekodi ya mgonjwa yanavuja. Kinachovuja ni membership — ukweli wa kuwa ndani — ambao ni hatari kwa sababu tofauti, si kwa sababu rekodi nzima imetolewa.
- **Haipunguzi** tofauti hizi hadi namba moja ya kichwa cha habari. Matokeo yenye nguvu na yanayoji rudia ni *pattern* — vipimo za jumla zinapunguza hatari halisi ya mtu mmoja, na wagonjwa waliowakilishwa kidogo hubeba sehemu kubwa ya hatari hiyo — katika hifadhidata na mamia ya modeli.

Ushahidi una nguvu kiasi gani?

Haya ni matokeo ya empirical na methodological, na yana nguvu. Pattern — vipimo za jumla za faragha zinapunguza kwa utaratibu hatari ya mtu mmoja, hatari iliyobaki inajikusanya katika makundi yaliyowakilishwa kidogo, na inazidi kuwa mbaya kadiri modeli capacity inavyoongezeka — ilionekana katika **hifadhidata saba za aina tofauti za data** na idadi kubwa ya modeli kwa kila hifadhidata. Upana huo ndio unaofanya dai la kipimo liwe la kuaminika badala ya artifact ya hifadhidata moja.

Tahadhari mbili za uaminifu. Kwanza, huu ni uthibitisho wa *hatari*, uliopimwa kwa kuendesha mashambulio ambayo waandishi wenyewe walijenga; unaonyesha mshambulizi mwenye uwezo *angeweza* kufanya vizuri kiasi gani chini ya dhana zilizotajwa, si mashambulio ya kweli hutokea mara ngapi. Pili, namba zinazovutia — mafanikio karibu kamili kwa wagonjwa fulani — zinaelezea watu walio katika hali mbaya zaidi *kwa makusudi*; hiyo ndiyo hoja, na zinapaswa kusomwa kama “mkia wa hatari ni mzito zaidi kuliko wastani unavyoonyesha,” si “wagonjwa wengi wako wazi.”

Msimamo unaofaa si alarm wala kupuuza: ni utafiti makini wa hifadhidata nyingi unaoonyesha kwamba njia ya kawaida ya kuthibitisha faragha ya AI ya kitabibu ni kipofu kwa kesi zake mbaya zaidi — na kesi hizo zinawaangukia wagonjwa wenye nafasi ndogo zaidi ya kubeba madhara.

Kwa nini ni muhimu

Mambo mawili yanafanya hii iwe zaidi ya footnote ya kiufundi.

Kwanza, inabadilisha maana ambayo “privacy-preserving” inapaswa kuruhusiwa kuwa nayo. Kama modeli inatolewa na alama ya wastani ya faragha, alama hiyo inaweza kuwa ndogo kwa kweli na bado kuficha kundi la wagonjwa wanaoweza kutambuliwa karibu kikamilifu kwa membership attack. Dai la vitendo la makala ni wazi: pima hatari ya faragha **kwa kila mgonjwa** kabla ya release, dhibiti nani anaweza kufikia modeli zilizowekwa kazini, na tumia mbinu kama **differential privacy** — kelele ndogo iliyokalibishwa kwa uangalifu wakati wa training ambayo hupunguza membership attacks, kwa gharama halisi na inayosimamiwa kwa usefulness ya modeli (privacy–utility trade-off ni wazi, si bure). “Tulipima wastani” isiendelee kuhesabika kama “tulikagua faragha.”

Pili, inaunganisha faragha na fairness. Makundi yale yale yasiyowakilishwa vya kutosha katika data za kitabibu — na hivyo tayari kuhudumiwa vibaya zaidi na AI ya kitabibu — yanaonekana kuwa yale ambayo faragha yao inalindwa kwa kiwango kidogo zaidi. Uwanja unaojitahidi kurekebisha ukosefu wa usawa wa kwanza hauwezi kuchukulia wa pili kama tatizo la idara nyingine. Ni watu wale wale.

Simulizi la kutia moyo kuhusu faragha ya AI ya kitabibu — *tulipima, wastani ni mdogo, tuko salama* — ndilo makala hii inalivunja. Si kuwafukuza watu kutoka teknolojia, bali kusogeza kiwango mahali kinapostahili: ahadi ambayo huwezi kudai umeitimiza mpaka umeikagua kwa mtu mwenye uwezekano mkubwa zaidi wa kuumizwa.

Muhtasari safi

Membership inference attacks hujaribu kujua kama rekodi ya mtu fulani ilikuwa ndani ya training data ya modeli ya AI — na kwa sababu membership yenyewe inaweza kufichua taarifa (kwa mfano kwamba ulikuwa na ugonjwa fulani), huo ni uvujaji halisi wa faragha hata kama rekodi yenyewe haionekani. Kazi za awali zilipima mara ngapi mashambulio hayo hufanikiwa *kwa wastani* katika hifadhidata. Utafiti huu ulipima *kwa kila mgonjwa*, katika hifadhidata saba za kitabibu (upigaji picha, ECG, electronic records) na modeli nyingi kwa kila moja, na ukapata kwamba vipimo za jumla zinapunguza hatari kwa kiasi kikubwa: baadhi ya watu wanaweza kutambuliwa karibu kikamilifu (attack AUC $\geq$ 0.95) hata wakati wastani wa hifadhidata unaonekana salama; walio hatarini zaidi ni kwa utaratibu wale wa makundi yaliyowakilishwa kidogo (makabila madogo, magonjwa nadra, upigaji picha isiyo ya kawaida); na tatizo huongezeka kadiri modeli zinavyokuwa kubwa. Waandishi hawasemi AI ya kitabibu iachwe — wanasema faragha ipimwe kwa mtu mmoja mmoja, access kwa modeli idhibitiwe, na differential privacy itumike. Ujumbe: “private kwa wastani” si ahadi ya faragha, na inaposhindwa huwaangukia zaidi wagonjwa ambao tayari wanalindwa kidogo.

Ukaguzi bila kupamba ukweli

Kile makala inaonyesha: Katika hifadhidata saba za kitabibu na modeli nyingi kwa kila moja, hatari ya membership inference kwa kila mgonjwa ni kubwa zaidi kwa baadhi ya watu kuliko vipimo za jumla zinavyoonyesha — hadi mafanikio karibu kamili ya shambulio (AUC $\geq$ 0.95) — na hatari hiyo iliyobaki inawaangukia kwa kiwango kisicho sawa makundi yaliyowakilishwa kidogo na huongezeka pamoja na uwezo wa modeli.

Kile kinachowezekana lakini hakijathibitishwa: Kwamba washambulizi wa dunia halisi tayari wanatumia hili dhidi ya modeli za kliniki zilizowekwa kazini; utafiti unaonyesha *uwezo na usambazaji wake* chini ya dhana zilizotajwa, si mara ngapi mashambulio hutokea nje ya maabara.

Kile haionyeshi: Kwamba AI ya kitabibu iachwe; kwamba kila modeli inavuja au modeli fulani tayari imevunjwa; kwamba *maudhui* ya rekodi ya mgonjwa, badala ya membership, yanafichuliwa; au disparity moja inayoweza kufupishwa kwa namba moja — matokeo thabiti ni pattern, si namba moja.

Mipaka mikuu: Inapima worst-case hatari kupitia mashambulio ya waandishi wenyewe, si incidents zilizoonekana; namba kali zinaeleza watu walio wazi zaidi kwa makusudi; na tofauti kamili kati ya makundi zinaonekana kama pattern thabiti katika hifadhidata badala ya kupunguzwa kuwa namba moja.

Msomaji wa kawaida anapaswa kuwa na uhakika kiasi gani? Uhakika mkubwa kwamba vipimo za jumla za faragha zinapunguza hatari ya mtu mmoja, kwamba hatari iliyobaki inajikusanya kwa wagonjwa waliowakilishwa kidogo, na kwamba hali hiyo huwa mbaya zaidi modeli size inapoongezeka — imeonyeshwa katika hifadhidata na modeli nyingi. Uhakika wa wastani kuhusu mara ngapi mashambulio haya hutokea katika dunia halisi, jambo ambalo utafiti haukupima. Usomaji salama si “AI ya kitabibu inavuja data yako,” wala “faragha imetatuliwa,” bali “ukaguzi wa kawaida wa faragha hauoni kesi zake mbaya zaidi, na kesi hizo zinawaangukia walio hatarini zaidi — hivyo ukaguzi wenyewe lazima ubadilike.”

Vyanzo

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