The Clean Paper

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

Egy orvosi MI átlagosan lehet privát, miközben egyes betegeket mégis felfed — leginkább az alulreprezentáltakat

Egy „adatvédelmet megőrzőnek" nevezett orvosi MI-modell minősítése rendszerint egyetlen átlagos számon nyugszik. Ez a tanulmány amellett érvel, hogy ez rossz teszt. A membership inference támadásokat vizsgálva — amelyek felfedhetik, hogy egy konkrét személy rekordja szerepelt-e a modell tanítóadatai között, és ezzel akár azt is elárulhatják, hogy bizonyos betegsége volt — a szerzők hét orvosi adathalmazon (képalkotás, ECG, elektronikus nyilvántartások) és számos modellen betegenként, nem összesítve mérték a kockázatot. A minta: az átlag alapján biztonságosnak tűnő modellek egyes személyeket szinte tökéletesen azonosíthatóvá tehetnek (attack AUC ≥ 0,95); a kitettek rendszeresen alulreprezentált csoportokból kerülnek ki; és a probléma a modellek növekedésével rosszabbodik. A szerzők nem az orvosi MI elhagyását kérik, hanem betegenkénti adatvédelmi mérést, a modellhozzáférés szabályozását és differenciális adatvédelmet. A kellemetlen lényeg: az „átlagosan privát" nem adatvédelmi garancia, és éppen azokat a betegeket hagyja cserben, akik eleve a legkevésbé védettek.

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

Az adatvédelmi garancia egy embernek tett ígéret, nem egy átlagnak

Amikor egy orvosi MI-modellt „adatvédelmet megőrzőnek” neveznek, ez az állítás rendszerint egyetlen számon alapul: a modell betanításához felhasznált összes betegre nézve alacsony annak átlagos esélye, hogy ki lehessen következtetni, szerepelt-e egy adott személy a tanítóhalmazban. Ez megnyugtatóan hangzik. A tanulmány halk, kellemetlen állítása azonban az, hogy rossz számot nézünk.

Az adatvédelem nem átlag. Minden egyes embernek tett ígéret arról, hogy a tanítóhalmazban való részvétele később nem fogja őt felfedni. Egy átlag pedig szinte mindenkinél betarthatja ezt az ígéretet, miközben néhány embernél teljesen megszegi. A kutatók betegenként mérték az adatvédelmi kockázatot, korábban nem használt felbontásban, és pontosan ezt találták: az összesített mutatók alapján biztonságosnak tűnő modellek bizonyos személyek tanítóhalmaz-tagságát szinte tökéletesen kiszivárogtathatják — és az érintettek aránytalanul gyakran éppen azok, akik eleve a legkevésbé védettek.

Mit csináltak a szerzők

A Moritz Knolle vezette csoport — Daniel Rückerttel (mindketten a Müncheni Műszaki Egyetemről), Georg Kaississzal (Hasso Plattner Institute), valamint az Imperial College London munkatársaival — egy jól ismert adatvédelmi fenyegetést, az úgynevezett **membership inference attackot**, vagyis tanítóhalmaz-tagságot következtető támadást vizsgálta meg más kérdésfeltevéssel. Nem azt kérdezték, hogy „az egész adathalmazon átlagosan milyen gyakran sikerül a támadás?”, hanem azt: „*ennél a konkrét betegnél* mekkora a kitettség?”

Ezt a betegenkénti elemzést **hét bevett orvosi adathalmazon** futtatták le, nagyon különböző adattípusokon — orvosi képeken, elektrokardiogramokon és elektronikus egészségügyi nyilvántartásokon —, adathalmazonként 200 modellen. Minden beteg és minden modell esetében megbecsülték, milyen magabiztosan tudná egy támadó eldönteni, hogy az illető adata része volt-e a tanítóhalmaznak, majd az eredményeket csoportok szerint is elemezték: betegségállapot, önbevallott rassz, biztosítás, nem és képalkotási protokoll alapján.

> **Mi is valójában a membership inference támadás?**
>
> Az itt vizsgált szivárgás finomabb annál, hogy „a modell kiköpi a kartonodat”. A *tagságról* szól — pusztán arról a tényről, hogy az adatod benne volt a tanítóhalmazban.
> Induljunk ki abból, amit egy ilyen modell normálisan csinál. Beadunk neki például egy mellkasröntgent, és valószínűségeket ad vissza: 78% esély tüdőgyulladásra, valamint értékeket szívmegnagyobbodásra, ödémára, konszolidációra. A támadás *kizárólag* ezt a hétköznapi diagnosztikai választ használja. Semmi különlegeset.
> A kihasznált gyengeség az, hogy a modell rendszerint kissé **magabiztosabb** pontosan azokkal

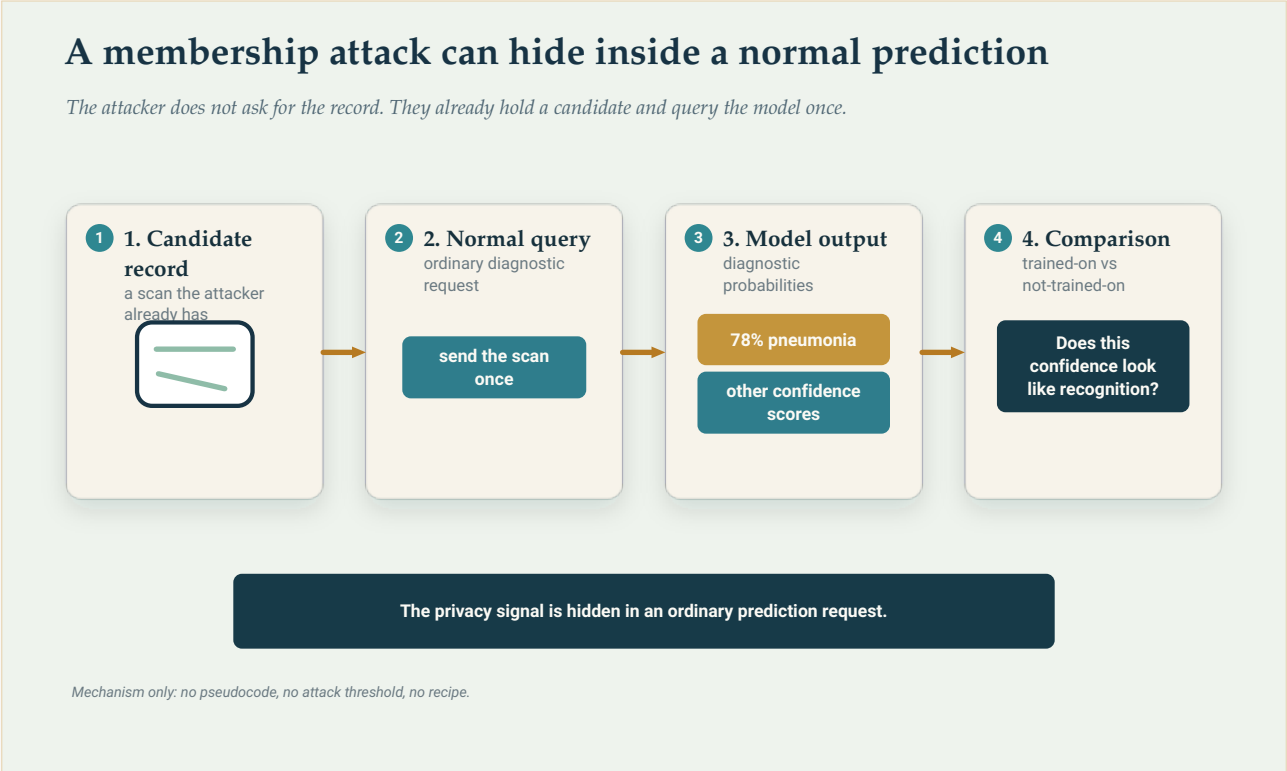
a példákkal kapcsolatban, amelyeken tanították, mint azokkal, amelyeket még soha nem látott. Képzeljünk el egy diákot, aki titokban előre látta a vizsgát: a már begyakorolt kérdésekre egy kicsit túl gyorsan, túl biztosan válaszol. A „membership inference” azt jelenti, hogy ebből az árulkodó jelből próbáljuk megállapítani, egy adott rekord a modell által látott példák között volt-e.

A támadás tehát lépésről lépésre így néz ki. Valaki tudni akarja, hogy egy adott személy felvételét felhasználták-e a modell tanításához. **Nem** magát a rekordot kéri a modelltől — már rendelkezik egy jelölt felvétellel, az illető saját képével vagy annak közeli másolatával. Egyszer elküldi szokásos diagnosztikai kérésként, és feljegyzi, mennyire magabiztos a válasz. Ezután megnézi, ez a magabiztosság inkább olyan modellre hasonlít-e, amely *tanult* ezen a felvételen, vagy olyanra, amely *nem*. Az összehasonlítást olcsón megteheti egy saját helyettesítő („referencia-”) modell betanításával, amely megtanulja, hogyan néz ki ez az árulkodó túlzott magabiztosság; ehhez egy hétköznapi számítógép is elég, különleges hardver nélkül. Ha a valódi modell szokatlanul biztos ebben a felvételben — mintha felismerné —, az erős bizonyíték arra, hogy a kép a tanítóadatok között volt.

A nyugtalanító rész az, hogy magán a kérésen semmi nem látszik támadásnak: ugyanaz a diagnosztikai lekérdezés, amelyet egy klinikus is használna, az adatvédelmi jel pedig egy hétköznapi előrejelzésben rejtőzik. Ettől konkrét a kockázat — még akkor is, ha eddig megadott laboratóriumi feltételek mellett demonstrálták, és nem valós támadásként figyelték meg.

Miért számít a tagság, ha maga a karton soha nem szivárog ki? Mert *a tagság önmagában tény*. Ha egy modellt például egy bizonyos daganatos immunterápiát kapó betegek adatain tanítottak, annak megerősítése, hogy a te rekordod benne van, elárulhatja, hogy valószínűleg ilyen daganatod volt — olyasmit, amit egy biztosítónak vagy munkáltatónak soha nem volna szabad kikövetkeztetnie. A tartalom zárva marad; a valahová tartozás ténye szökik ki.

A szerzők világosan rögzítik a feltételeket — hozzáférés a modell szokásos előrejelzéseihez, egy jelölt rekord és a támadó saját referenciamodellje —, nem használati útmutatóként, hanem azért, hogy a fenyegetést konkrétan lehessen mérlegelni.



A támadáshoz nincs szükség külön adatvédelmi API-ra. Egy normál diagnosztikai lekérdezés is szivárogtathat tagsági jelet, ha a modell szokatlanul magabiztos egy korábban látott rekord esetében.

Mit találtak

Három eredményt, amelyek egymás után egyre élesebbek.

Az átlagok elrejtik a kitetteket. Összesítve — a szokásos módon — sok modell megnyugtatóan biztonságosnak látszik: a támadás a betegek többségénél nem jobb a véletlennél. A betegenkénti nézet egészen más képet adott. Knolle a tanulmány bejelentésében úgy fogalmazott, hogy a korábbi értékelések csak az összes beteg átlagos kockázatát mérték, míg ők először vizsgálták azt az egyes betegek szintjén — és az eredmény egészen más. Egyes személyeknél a támadás **szinte tökéletesen** működött: az attack AUC **0,95 vagy magasabb** volt (0,5 a pénzfeldobás, 1,0 a hibátlan találat), miközben az egész adathalmazra számolt átlag nem látszott jobbnak a véletlennél.

Az átlag a véletlen szintje közelében lehet, miközben a rekordok egy kis szélső csoportja sokkal könnyebben azonosítható. A tanulmány lényege, hogy az adatvédelmet betegenként is ellenőrizni kell, nem csak összesítve.

A kitettség egyenlőtlen — és az alulreprezentáltakat sújtja. A szinte tökéletes támadásnak leginkább kitett betegek rendszeresen az adatban **alulreprezentált csoportokból** kerültek ki: etnikai kisebbségekből, ritka betegségfenotípusokkal vagy szokatlan képalkotási jellemzőkkel. Az elektronikus egészségügyi adatbázisban a fekete betegek a vártnál **31%-kal gyakrabban** jelentek meg a legsérülékenyebb rekordok között; a mammográfiás adathalmazban pedig a rosszindulatúság szempontjából gyanúsnak jelölt felvételek megdöbbentő **+1179%-os** felülreprezentációt mutattak ebben a veszélyzónában. (Ez a két szám példa, nem a teljes kép; a tanulmány több csoportnál is hasonló torzulást talál. Ráadásul *relatív* felülreprezentációról van szó a kis, extrém kockázatú szélső csoporton belül: arról, hogy ezek a betegek mennyivel gyakrabban fordulnak elő a legkitettebb rekordok között, nem arról, hogy a fekete vagy gyanús mammográfiás betegek mekkora hányada kitett.) A modellnek kevesebb hasonló példája van, amelyek között ezek a betegek „elmosódhatnának", így rekordjaik jobban kilógnak — márpedig a membership inference pontosan ezt a kilógást érzékeli. Az adatvédelmi kudarc tehát nem véletlenszerű; a már eleve peremhelyzetben lévőkre koncentrálódik.

A nagyobb modellek rontanak a helyzeten. A szinte tökéletes támadásoknak kitett betegek száma meredeken nőtt a modellkapacitással. Egy bőrgyógyászati adathalmazon az arány a legkisebb modell gyakorlatilag nulla értékéről a legnagyobbnál körülbelül **minden tizedik betegre** emelkedett. Ahogy az orvosi MI-modellek egyre nagyobbak és erősebbek lesznek — és a terület éppen ebbe az irányba halad —, ez a konkrét kockázat a szerzők szerint súlyosbodik, nem enyhül.

Rückert összegzése tömör: „Ez nem elfogadható kockázat. Az egészségügyi adatok rendkívül érzékenyek."

Miért rossz teszt az, hogy „átlagosan biztonságos”

Könnyű egy alacsony átlagos kockázatot úgy olvasni, mint egy tiszta egészségügyi bizonyítványt. A tanulmány központi tanulsága azonban az, hogy az átlagolás itt nem csupán pontatlan — rossz dolgot mér.

Egy adatvédelmi garancia csak akkor jelent valamit, ha a legnagyobb kockázatnak kitett emberre is érvényes, nem csak a középen lévőre. Ha 1000 betegből 999 gyakorlatilag azonosíthatatlan, de egyet szinte tökéletesen fel lehet ismerni, akkor a modell nem „99,9%-ban privát” semmilyen, az érintett egyetlen ember számára értelmes értelemben. Ha pedig ez az egy ember előre jelezhető módon ritka betegségben szenved vagy etnikai kisebbséghez tartozik, a mérőszám nem egyszerűen hiányos: csendben diszkriminatív. A többség biztonságát méri, majd mindenki biztonságának nevezi.

Ezt a nézőpontváltást kényszeríti ki a tanulmány: a *mennyire privát ez a modell átlagosan?* kérdésről arra, hogy *ki a leginkább kitett beteg, és ki ő?* Ez két külön kérdés, és adatvédelmi kérdés valójában csak a második.

Mit nem bizonyít — és nem is állít — a tanulmány

- **Nem** azt mondja, hogy fel kell hagyni az orvosi MI-vel. A szerzők a kockázat mérsékléséről beszélnek, nem visszavonulásról: mérjük és javítsuk a problémát, ne álljunk le a fejlesztéssel.
- **Nem** jelenti azt, hogy minden orvosi modell szivárog, vagy hogy bármely konkrét, élesben használt modellt már megtámadtak. Azt mutatja, hogy *a kockázat létezik és egyenlőtlenül oszlik meg*, miközben a szokásos összesített mutatók elrejtik.
- **Nem** jelenti azt, hogy az egészségügyi adataid már most ki vannak téve. A támadáshoz meghatározott feltételek kellenek — hozzáférés a modellhez, egy jelölt rekord és a támadó saját infrastruktúrája —, nem alkalmi képességről van szó.
- **Nem** azt mutatja, hogy a betegadat *tartalma* szivárog. A tagság ténye szivárog ki, ami más okból veszélyes, nem azért, mert a teljes rekord megjelenik.
- **Nem** redukálja az eltéréseket egyetlen látványos számra. A robusztus, megismétlődő eredmény maga a *minta*: az összesített mutatók alábecsülik az egyéni kockázatot, és az alulreprezentált betegek viselik ennek legnagyobb részét — több adathalmazon és több száz modellen át.

Mennyire erős a bizonyíték?

Ez empirikus, módszertani eredmény, és robusztus. Az a minta, hogy az összesített adatvédelmi mutatók rendszeresen alábecsülik a betegenkénti kockázatot, a maradék kockázat az alulreprezentált csoportokra koncentrálódik, és a modellkapacitással nő, **hét, különböző adattípusú adathalmazon** és adathalmazonként sok modellen is megmaradt. Éppen ez a szélesség teszi a mérési állítást hihetővé egyetlen adathalmaz különcsége helyett.

Két fontos fenntartás van. Először: ez a *kockázat* demonstrációja, amelyet a szerzők saját támadásaik futtatásával mértek. Azt jellemzi, mire *lehetne* képes egy megfelelő támadó a megadott feltételek között, nem azt, milyen gyakran történik ilyen támadás a valóságban. Másodszor: a látványos számok — egyes betegeknél a szinte tökéletes támadási siker — szándékosan a legrosszabb helyzetben lévő személyeket írják le. Ez maga a vizsgálat lényege; úgy kell olvasni, hogy „az eloszlás széle sokkal súlyosabb, mint amit az átlag sugall", nem úgy, hogy „a betegek többsége ki van téve".

A megfelelő hozzáállás tehát sem a riadalom, sem a legyintés: ez egy gondos, több adathalmazra kiterjedő tanulmány, amely megmutatja, hogy az orvosi MI adatvédelmének szokásos minősítése vak a saját legrosszabb eseteire — és ezek az esetek éppen azokat sújtják, akiknek a legkevesebb tartalékuk van.

Miért fontos

Két dolog emeli ezt technikai lábjegyzet fölé.

Először is megváltoztatja, mit szabadna jelentenie annak, hogy „adatvédelmet megőrző". Ha egy modellt átlagos adatvédelmi pontszámmal adnak ki, ez a pontszám lehet valóban alacsony, és mégis elrejthet egy betegcsoportot, amelyet membership inference támadással szinte tökéletesen azonosítani lehet. A tanulmány gyakorlati követelése konkrét: kiadás előtt mérjük az adatvédelmi kockázatot **betegenként**, szabályozzuk, ki férhet hozzá az élesben futó modellekhez, és alkalmazzunk olyan módszereket, mint a **differenciális adatvédelem** — a tanítás során hozzáadott kis, gondosan kalibrált zajt, amely gyengíti a tagsági támadásokat, valós és kezelendő ár mellett a modell hasznosságában. A privacy–utility kompromisszum létezik; nincs ingyenes megoldás. Az, hogy „ellenőriztük az átlagot", nem számíthat többé teljes ellenőrzésnek.

Másodszor összefonja az adatvédelmet a méltányossággal. Ugyanazok a csoportok, amelyek alulreprezentáltak az orvosi adatokban — és ezért már eleve rosszabb kiszolgálást kaphatnak az orvosi MI-től —, azok, akiknek az adatvédelmét ugyanez az MI a legkevésbé védi. Egy terület, amely keményen dolgozik az első egyenlőtlenségen, nem kezelheti a másodikat másvalaki problémájaként. Ugyanazokról az emberekről van szó.

Az orvosi MI adatvédelméről szóló megnyugtató történet — *megmértük, alacsony az átlag, minden rendben* — az, amit ez a tanulmány szétszed. Nem azért, hogy bárkit elriasszon a technológiától, hanem hogy oda tegye a mércét, ahová való: egy ígéretet csak akkor állíthatunk betartottnak, ha annál az embernél is ellenőriztük, aki a leginkább sérülhet.

Röviden

A membership inference támadások azt próbálják megállapítani, hogy egy konkrét személy rekordja szerepelt-e egy MI-modell tanítóadatai között. Mivel maga a tagság is árulkodó lehet — például felfedheti, hogy valaki egy bizonyos betegségben szenvedett —, ez valódi adatvédelmi szivárgás akkor is,

ha maga a rekord soha nem kerül elő. Korábbi munkák azt mérték, milyen gyakran sikerülnek az ilyen támadások *átlagosan* egy adathalmazon. Ez a tanulmány *betegenként* mérte a kockázatot hét orvosi adathalmazon — képalkotás, ECG, elektronikus egészségügyi nyilvántartások — és sok modellen. Az eredmény: az összesített mutatók erősen alábecsülhetik a kockázatot; egyes személyek szinte tökéletesen azonosíthatók (attack AUC $\geq 0{,}95$) akkor is, amikor az egész adathalmaz átlaga biztonságosnak tűnik; a legsérülékenyebbek rendszeresen az alulreprezentált csoportok tagjai; és a probléma a modellmérettel nő. A szerzők nem az orvosi MI elhagyását javasolják, hanem az egyéni szintű adatvédelmi mérést, a modellhozzáférés szabályozását és a differenciális adatvédelmet. A tanulság: az „átlagosan privát" nem adatvédelmi garancia, és éppen azoknál mond csődöt, akik eleve a legkevésbé védettek.

Tárgyilagos mérleg

Mit mutat a tanulmány: Hét orvosi adathalmazon és sok modellen a betegenkénti membership-inference kockázat egyes személyeknél jóval magasabb annál, amit az összesített mutatók sugallnak — egészen a szinte tökéletes támadási sikerig (AUC $\geq 0{,}95$) —, és ez a maradék kockázat aránytalanul az alulreprezentált csoportokra esik, valamint nő a modellkapacitással.

Mi hihető, de nincs bizonyítva: Hogy valós támadók már most használják ezt éles klinikai modellek ellen. A tanulmány a *képességet és annak eloszlását* demonstrálja megadott feltételek mellett, nem a valós támadások gyakoriságát.

Mit nem mutat: Hogy fel kell hagyni az orvosi MI-vel; hogy minden modell szivárog vagy bármely konkrét éles modell sérült; hogy a betegkarton *tartalma* (és nem a tagság) kerül ki; egyetlen, mindent összefoglaló eltérési számot — a robusztus eredmény a minta, nem egyetlen érték.

Fő korlátok: A szerzők saját támadásaival mért legrosszabb eseti kockázatról van szó, nem megfigyelt incidensekről; a drámai számok szándékosan a leginkább kitett személyeket írják le; a csoportok közötti pontos különbségek több adathalazon következetes mintaként jelennek meg, nem egyetlen számmá sűrítve.

Mennyire bízhat benne egy általános olvasó? Nagy bizalommal abban, hogy az összesített adatvédelmi mutatók alábecsülik az egyéni kockázatot, hogy a maradék kockázat az alulreprezentált betegekre koncentrálódik, és hogy ez a modellmérettel romlik — sok adathalmazon és modellen demonstrálták. Közepes bizalommal a valós támadások gyakoriságában, mert azt ez a tanulmány nem méri. A helyes olvasat nem az, hogy „az orvosi MI kiszivárogtatja az adataidat", és nem is az, hogy „az adatvédelem meg van oldva", hanem az: „a szokásos adatvédelmi ellenőrzés nem látja a saját legrosszabb eseteit, amelyek a legsérülékenyebb betegeket sújtják — ezért az ellenőrzést meg kell változtatni."

Források

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