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

Lekárska AI môže byť v priemere súkromná, no napriek tomu odhaľovať konkrétnych pacientov — najmä tých nedostatočne zastúpených

Tvrdenie, že medicínsky model AI „chráni súkromie“, zvyčajne stojí na jedinom priemernom čísle. Táto štúdia tvrdí, že ide o nesprávny test. Pri skúmaní útokov na odvodenie členstva — ktoré zisťujú, či bol záznam konkrétnej osoby v trénovacích dátach modelu, a môžu tak prezradiť, že mala určitú chorobu — autori merali riziko po jednotlivých pacientoch, nie iba súhrnne, naprieč siedmimi súbormi medicínskych dát a mnohými modelmi. Výsledný vzorec: modely, ktoré vyzerajú v priemere bezpečne, stále umožňujú útočníkovi takmer dokonale identifikovať niektoré osoby (AUC útoku ≥ 0,95); ohrození sú systematicky ľudia z nedostatočne zastúpených skupín, napríklad príslušníci etnických menšín, pacienti so zriedkavými chorobami či nezvyčajnými snímkami; a s veľkosťou modelov sa problém zhoršuje. Autori nenavrhujú opustiť medicínsku AI — žiadajú merať súkromie na úrovni pacienta, kontrolovať prístup k modelom a používať diferenciálne súkromie. Nepríjemné jadro výsledku znie: „súkromné v priemere“ nie je zárukou súkromia a zlyháva práve u pacientov, ktorí sú už najmenej chránení.

Written by Lucio Vaglio · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso · 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

Záruka súkromia je sľub človeku, nie priemernému človeku

Keď sa o modeli medicínskej AI povie, že „chráni súkromie“, tvrdenie sa zvyčajne opiera o jediné číslo: priemerná pravdepodobnosť, že sa dá určiť členstvo niektorého z pacientov, na ktorých údajoch sa model trénoval, je nízka. Znie to upokojujúco. Nenápadným a nepríjemným záverom tejto práce je, že ide o nesprávne číslo.

Súkromie nie je priemer. Je to sľub daný každému jednotlivcovi, že jeho účasť v tréningovom súbore ho neskôr neodhalí. Priemer pritom môže tento sľub splniť takmer každému a niekoľkým ľuďom ho úplne porušiť. Výskumníci merali riziko ochrany súkromia osobitne pre každého pacienta, s doteraz nepoužitým rozlíšením, a zistili presne toto: modely, ktoré v súhrne vyzerajú bezpečne, môžu takmer dokonale prezradiť členstvo konkrétnych ľudí. A medzi odhalenými sú neúmerne často tí, ktorí už patria k najmenej chráneným.

Čo autori urobili

Tím vedený Moritzom Knollem, s Danielom Rückertom (obaja z Technickej univerzity v Mníchove), Georgom Kaissisom (Hasso-Plattner-Institut) a kolegami z Imperial College London vzal známu hrozbu pre súkromie, **útok na odvodenie členstva**, a zmenil otázku. Namiesto „ako často je tento útok v priemere úspešný v celom súbore dát?“ sa pýtal: „Ako veľmi je vystavený *tento konkrétny pacient*?“

Analýzu na úrovni pacientov vykonali v **siedmich zavedených medicínskych súboroch dát**, ktoré pokrývali veľmi odlišné typy údajov — medicínske snímky, elektrokardiogramy a elektronické zdravotné záznamy — a naprieč nimi použili po 200 modelov. Pri každom pacientovi a modeli odhadli, s akou istotou môže útočník určiť, či bol pacientov záznam súčasťou tréningových dát. Výsledky potom rozdelili podľa skupín: zdravotný stav, rasa uvedená samotným pacientom, poistenie, pohlavie a zobrazovací protokol.

Čo vlastne je útok na odvodenie členstva

Únik je tu menej priamočiary než „model vypľuje váš záznam“. Ide o *členstvo* — samotný fakt, že vaše údaje boli v tréningovom súbore.

Začnime tým, čo taký model bežne robí. Zadáte mu snímku, napríklad röntgen hrudníka, a model vráti pravdepodobnosti: 78 % pravdepodobnosť zápalu pľúc spolu s hodnoteniami kardiomegálie, edému a konsolidácie. Táto bežná diagnostická odpoveď je *jediné*, čo útok používa. Nič nezvyčajné.

Využíva túto slabinu: model si je zvyčajne o niečo **istejší** pri presných príkladoch, na ktorých sa trénoval, než pri tých, ktoré nikdy nevidel. Predstavte si študenta, ktorý tajne videl skúšku vopred — pri otázkach, ktoré si precvičil, prichádzajú odpovede trochu prirýchlo a s trochu priveľkou istotou. Určiť podľa takéhoto signálu, či konkrétny záznam patril medzi príklady, ktoré model videl, znamená „odvodiť členstvo“.

Útok teda prebieha takto. Niekto chce vedieť, či sa snímka konkrétnej osoby použila pri trénovaní modelu. Od modelu si záznam nepýta — kandidátsku snímku už má, či už originál, alebo veľmi podobnú kópiu. Raz ju odošle ako bežnú diagnostickú požiadavku a zaznamená istotu odpovede. Potom overí, či táto istota vyzerá skôr ako pri modeli, ktorý sa na snímke *trénoval*, alebo ako pri modeli, ktorý ju *nevidel*. Porovnanie môže lacno vytvoriť natrénovaním vlastného zástupného „referenčného modelu“ na bežnom počítači bez osobitného hardvéru. Ak je skutočný model pri snímke nezvyčajne istý, akoby ju spoznával, je to silný dôkaz, že bola v tréningových dátach.

Znepokojujúce je, že požiadavka vôbec nevyzerá ako útok: ide o rovnaký diagnostický dotaz, aký by zadal klinik, a signál súvisiaci so súkromím sa skrýva v bežnej predikcii. Práve preto je riziko konkrétne — hoci sa zatiaľ preukázalo za stanovených laboratórnych predpokladov, nie pri zachytenom útoku v reálnom prostredí.

Prečo na členstve záleží, ak samotný záznam nikdy neunikne? Pretože *členstvo je fakt*. Ak sa model trénoval na pacientoch, ktorí dostali určitú imunoterapiu proti rakovine, potvrdenie, že váš záznam do súboru patrí, prezrádza, že ste túto rakovinu pravdepodobne mali. Poisťovňa ani zamestnávateľ by taký údaj nikdy nemali vedieť odvodiť. Obsah zostáva utajený; uniká fakt príslušnosti.

Autori tieto predpoklady jasne stanovili — prístup k bežným predpovediam modelu, kandidátskemu záznamu a vlastnému referenčnému modelu útočníka — nie ako návod, ale aby sa hrozba dala rozumne posúdiť a nie len prehliadnuť.

Útok na odvodenie členstva sa môže skrývať v bežnej predikcii

Útočník nežiada o záznam. Už má kandidátny záznam a model sa opýta raz.

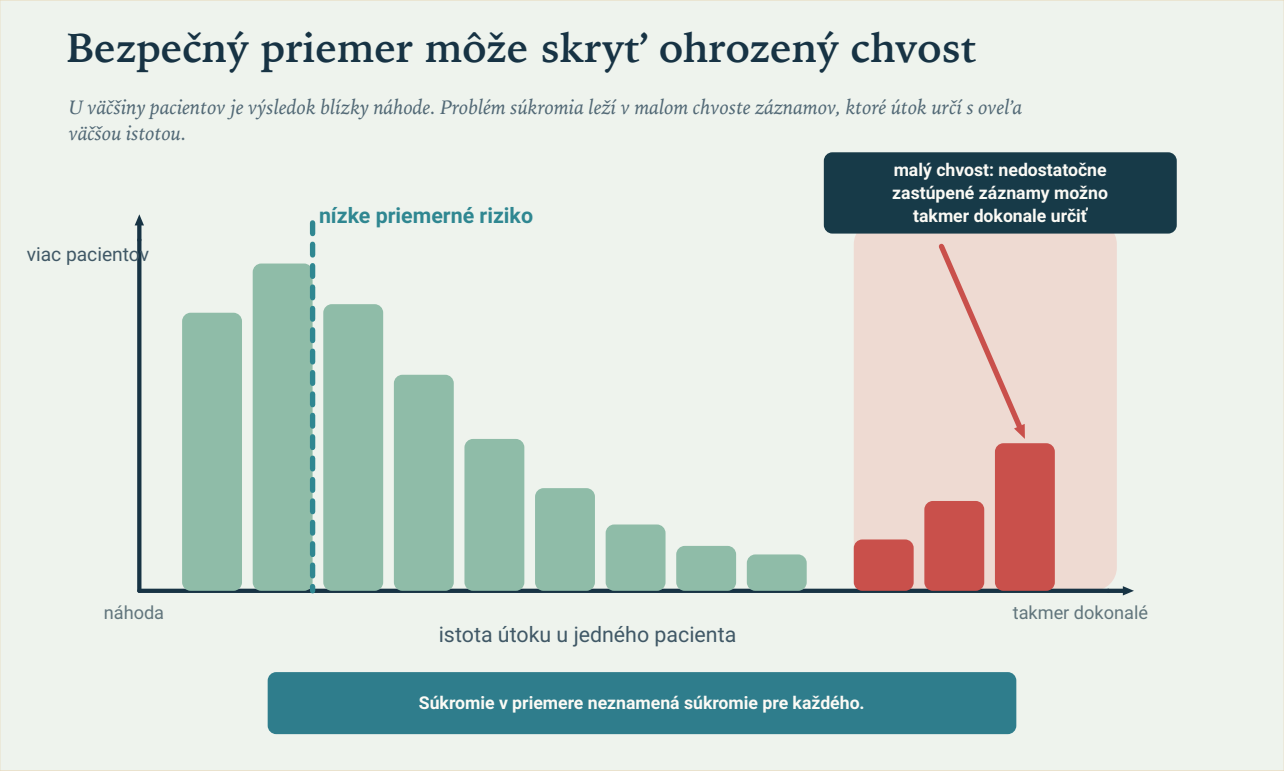
Len princíp: bez pseudokódu, prahu útoku či návodu.

Útok nepotrebuje osobitné rozhranie API pre súkromie. Bežný diagnostický dotaz môže prezradiť signál členstva, ak je model pri zázname, ktorý už videl, nezvyčajne istý.

Čo našli

Tri zistenia, každé ostrejšie než to predchádzajúce.

Priemery skrývajú odhalených. Ak sa to meria v súhrne — bežným spôsobom — mnohé z týchto modelov pôsobia upokojujúco súkromne: útok pre väčšinu pacientov nie je lepší než náhoda. Pohľad na pacienta hovoril iný príbeh. Ako Knolle uviedol v oznámení štúdie, predchádzajúce hodnotenia "merali len priemerné riziko u všetkých pacientov. Prvýkrát sme skúmali riziko na úrovni jednotlivých pacientov — a obraz je úplne iný." Pre niektorých jednotlivcov útok uspeje **takmer dokonale** — autori ho merajú ako AUC útoku **0,95 alebo viac** (0,5 je hod mincou, 1,0 je bezchybné) — aj keď priemer v celej dátovej sade nevyzeral lepšie než náhoda.



Priemer môže zostať blízko náhodnej úspešnosti, zatiaľ čo malý chvost záznamov je oveľa viac vystavený. Práca ukazuje, že súkromie treba kontrolovať na úrovni pacientov, nielen súhrnne.

Vystavenie je nerovnomerné — a dopadá na nedostatočne zastúpených. Najzraniteľnejší voči takmer dokonale úspešnému útoku boli systematicky pacienti zo skupín **nedostatočne zastúpených v údajoch**: príslušníci etnických menšín, ľudia s fenotypmi zriedkavých ochorení a s nezvyčajnými zobrazovacími znakmi. V súbore elektronických záznamov sa černošskí pacienti medzi najzraniteľnejšími záznamami objavovali **o 31 % častejšie**, než sa očakávalo; v mamografickom súbore boli snímky označené ako podozrivé z malignity v tejto nebezpečnej zóne nadmerne zastúpené o výrazných **+1**

179 %. (Tieto dve čísla sú príklady, nie celý obraz — práca uvádza rovnaké skreslenie pri viacerých skupinách — a ide o *relatívne* nadmerné zastúpenie v malom chvoste extrémneho rizika: o koľko častejšie sa títo pacienti objavujú medzi najvystavenejšími záznamami, nie aký podiel černošských pacientov alebo pacientov s podozrivou mamografiou je vystavený.) Model má menej podobných príkladov, medzi ktorými by sa títo pacienti stratili, takže ich záznamy vyčnievajú — a práve to útok na odvodenie členstva zachytáva. Zlyhanie súkromia nie je náhodné; sústreďuje sa na ľudí, ktorí už stoja na okraji.

Väčšie modely problém zhoršujú. Počet pacientov vystavených takmer dokonale úspešným útokom prudko rástol s kapacitou modelu — v jednom dermatologickom súbore dát sa podiel zvýšil z prakticky nuly pri najmenšom modeli na približne **jedného z desiatich** pri najväčšom. Keďže modely medicínskej AI rastú a zvyšujú svoje schopnosti, čo je smer celého odboru, toto konkrétne riziko sa podľa autorov zvyšuje, nie znižuje.

Rückertovo zhrnutie je priamočiare: “Toto nie je tolerovateľné riziko. Zdravotné údaje sú veľmi citlivé.”

Prečo je “bezpečný v priemere” nesprávny test

Je lákavé chápať nízke priemerné riziko ako potvrdenie bezpečnosti. Hlavným poučením práce je, že priemerovanie tu nie je iba nepresné — meria nesprávnu vec.

Záruka súkromia má zmysel iba vtedy, ak platí pre najohrozenejšiu osobu, nie pre človeka uprostred rozdelenia. Model, pri ktorom nemožno identifikovať 999 z 1 000 pacientov, ale jedného možno identifikovať takmer dokonale, nie je „na 99,9 % súkromný“ v žiadnom zmysle, ktorý by bol pre tohto človeka podstatný. Ak je ním navyše predvídateľne pacient so zriedkavým ochorením alebo príslušník etnickej menšiny, metrika nie je len neúplná, ale nenápadne diskriminačná. Oznamuje bezpečnosť väčšiny a vydáva ju za bezpečnosť všetkých.

Práca si vynucuje posun od otázky *nakoľko súkromný je tento model v priemere?* k otázke *ktorý pacient je najviac vystavený a kto to je?* Sú to odlišné otázky a iba druhá sa skutočne týka súkromia.

Čo toto *ne* dokazuje — ani netvrdí

- **Nehovorí**, že treba medicínsku AI opustiť. Autori hovoria o zmiernení rizika, nie o ústupe: riziko treba merať a napraviť, nie zastaviť vývoj.
- To **neznamená**, že každý medicínsky model uniká, alebo že niektorý nasadený model bol napadnutý. Ukazuje, že *riziko existuje a je nerovnomerne rozdelené*, a že štandardné agregované metriky ho prehliadajú.
- **Neznamená**, že vaše záznamy sú už odhalené. Útok vyžaduje osobitné podmienky — prístup k modelu, kandidátsky záznam a vlastnú infraštruktúru útočníka — nejde o bežnú schopnosť.
- **Neukazuje**, že uniká *obsah* pacientovho záznamu. Uniká členstvo, teda fakt zaradenia, ktorý je

nebezpečný z iného dôvodu, nie preto, že by model záznam priamo vydal.

- Neznižuje rozdiely na jedno hlavné číslo. Silným, reprodukovateľným výsledkom je *vzor* — agregované metriky podhodnotujú individuálne riziko a nedostatočne zastúpení pacienti ho znášajú najviac — naprieč dátovými súbormi a stovkami modelov.

Aké silné sú dôkazy?

Ide o empirický metodologický výsledok, ktorý je veľmi robustný. Vzorec — súhrnné metriky súkromia systematicky podhodnocujú riziko jednotlivých pacientov, zvyškové riziko sa sústreďuje na nedostatočne zastúpené skupiny a rastie s kapacitou modelu — sa udržal v **siedmich súboroch dát rôznych typov** a pri veľkom počte modelov v každom z nich. Práve táto šírka robí tvrdenie o meraní dôveryhodným, nie artefaktom jediného súboru dát.

Dve poctivé výhrady. Po prvé, ide o ukážku *rizika* meraného útokmi, ktoré vytvorili samotní autori; opisuje, ako úspešný by schopný útočník *mohol* byť za stanovených predpokladov, nie ako často k útokom dochádza v skutočnosti. Po druhé, výrazné čísla — takmer dokonalá úspešnosť útoku pri niektorých pacientoch — zámerne opisujú najhoršie postavených jednotlivcov. To je podstata analýzy a čísla treba čítať ako „chvost je oveľa ťažší, než naznačuje priemer“, nie ako „väčšina pacientov je vystavená“.

Vhodný postoj nie je ani poplach, ani odmietnutie: starostlivá štúdia z viacerých dátových súborov, ktorá ukáže, že štandardný spôsob certifikácie súkromia v medicínskej AI je slepý voči vlastným najhorším prípadom — a že tie najhoršie prípady dopadajú na pacientov s najmenším rezervným priestorom.

Prečo je to dôležité

Dve veci z toho robia viac než len technickú poznámku pod čiarou.

Po prvé, mení význam označenia „chráni súkromie“. Model možno zverejniť s naozaj nízkym priemerným skóre rizika, ktoré napriek tomu skrýva podskupinu pacientov takmer dokonale identifikovateľných útokom na odvodenie členstva. Praktická požiadavka práce je konkrétna: pred zverejnením hodnotiť riziko súkromia **pri každom pacientovi**, kontrolovať prístup k nasadeným modelom a používať metódy ako **diferenciálne súkromie** — malý, starostlivo kalibrovaný šum pridaný počas trénovania, ktorý tlmí útoky na členstvo za skutočnú a riadenú cenu v podobe užitočnosti modelu (kompromis medzi súkromím a užitočnosťou je výslovný, nie bezplatný). Veta „skontrolovali sme priemer“ by už nemala znamenať, že kontrola prebehla.

Po druhé, spája súkromie so spravodlivosťou. Tie isté skupiny, ktoré sú v medicínskych údajoch nedostatočne zastúpené — a teda už horšie obsluhované medicínskou AI — sú tie, ktorých súkromie AI chráni najmenej. Odbor, ktorý tvrdo pracuje na prvej nerovnosti, nemôže druhú považovať za oddelenie niekoho iného. Sú to tí istí ľudia.

Upokojujúci príbeh o súkromí v medicínskej AI — *merali sme to, priemer je nízky, sme v poriadku* — je príbeh, ktorý tento článok rozoberá na kúsky. Nie preto, aby niekoho odradil od technológie, ale aby sa štandard presunul tam, kam patrí: sľub, ktorý môžete dodržať len vtedy, ak ste ho skontrolovali pre osobu, ktorá je najpravdepodobnejšie zranená.

Čisté zhrnutie

Útoky na odvodenie členstva sa pokúšajú určiť, či bol záznam konkrétneho človeka v tréningových dátach modelu AI. Keďže už samotné členstvo môže prezradiť citlivý fakt, napríklad prekonanie určitej choroby, ide o skutočný únik súkromia, aj keď sa obsah záznamu nikdy neobjaví. Predchádzajúce práce merali, ako často sú takéto útoky úspešné *v priemere* v celom súbore dát. Táto štúdia merala riziko *pri každom pacientovi* v siedmich medicínskych súboroch dát (snímky, EKG a elektronické záznamy) a pri mnohých modeloch. Zistila, že súhrnné metriky riziko výrazne podhodnocujú: niektorých jednotlivcov možno identifikovať takmer dokonale (AUC útoku $\geq$ 0,95), aj keď celkový priemer vyzerá bezpečne; najzraniteľnejší sú systematicky ľudia z nedostatočne zastúpených skupín (etnické menšiny, zriedkavé ochorenia, nezvyčajné zobrazovacie znaky); a problém rastie s veľkosťou modelu. Autori nenavrhujú opustiť medicínsku AI — žiadajú merať súkromie na úrovni jednotlivcov, kontrolovať prístup k modelom a používať diferenciálne súkromie. Záver: „súkromné v priemere“ nie je záruka súkromia a zlyhanie zvyčajne dopadá na tých, ktorí sú už najmenej chránení.

Kontrola bez nezmyslov

Čo práca ukazuje: V siedmich medicínskych súboroch dát a pri mnohých modeloch je riziko odvodenia členstva pri niektorých pacientoch oveľa vyššie, než naznačujú súhrnné metriky — až po takmer dokonalú úspešnosť útoku (AUC $\geq$ 0,95). Toto zvyškové riziko neúmerne dopadá na nedostatočne zastúpené skupiny a rastie s kapacitou modelu.

Čo je pravdepodobné, ale nie dokázané: Že skutoční útočníci už túto slabinu využívajú proti nasadeným klinickým modelom; štúdia ukazuje *možnosti útoku a ich rozdelenie* za stanovených predpokladov, nie frekvenciu útokov v reálnom prostredí.

Čo neukazuje: Že medicínska AI by mala byť opustená; že každý model uniká alebo že bol narušený konkrétny nasadený model; že je zverejnený *obsah* záznamu pacienta (namiesto členstva); ani jediný kvantifikovaný údaj o rozdiele — robustný výsledok je vzor, nie jedno číslo.

Hlavné obmedzenia: Riziko najhoršieho prípadu sa meria pomocou útokov vytvorených autormi, nie podľa pozorovaných incidentov; výrazné čísla zámerne opisujú najvystavenejších jednotlivcov; a presné rozdiely medzi skupinami sa ukazujú ako konzistentný vzorec naprieč súbormi dát, nie ako jediné číslo.

Akú dôveru by mal mať bežný čitateľ? Vysokú v to, že súhrnné metriky súkromia podhodnocujú riziko jednotlivcov, že zvyškové riziko sa sústreďuje na nedostatočne zastúpených pacientov a že

rastie s veľkosťou modelu — výsledok sa preukázal v mnohých súboroch dát a modeloch. Strednú, pokiaľ ide o skutočnú frekvenciu takýchto útokov, ktorú štúdia nemeria. Bezpečné čítanie neznie ani „medicínska AI vypúšťa vaše údaje“, ani „súkromie je vyriešené“, ale: „štandardná kontrola súkromia nevidí vlastné najhoršie prípady a tie dopadajú na najzraniteľnejších pacientov, preto sa kontrola musí zmeniť“.

Zdroje

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-ai

Medical Xpress (2026) — coverage of the study
https://medicalxpress.com/news/2026-06-patient-groups-vulnerable-privacy-medical.html