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

Lékařská AI může být v průměru soukromá, a přesto odhalovat konkrétní pacienty — zejména ty nedostatečně zastoupené

Označení modelu lékařské AI za „chránící soukromí“ obvykle stojí na jediném průměrném čísle. Tato studie tvrdí, že jde o nesprávný test. Při zkoumání útoků na odvození členství — které odhalují, zda byl záznam konkrétní osoby v trénovacích datech modelu, a mohou tak prozradit, že dotyčný trpěl určitou nemocí — autoři měřili riziko pro každého pacienta zvlášť, nikoli pouze souhrnně. Analýza zahrnula sedm lékařských datových sad (snímky, EKG a elektronické záznamy) a u každé řadu modelů. Vzorec byl následující: modely, které v průměru vypadají bezpečně, mohou útočníkovi stále umožnit téměř dokonale určit konkrétní osoby (AUC útoku $\geq 0{,}95$); ohroženi jsou systematicky lidé z nedostatečně zastoupených skupin (etnické menšiny, vzácná onemocnění, neobvyklé snímky); a s růstem modelů se problém zhoršuje. Autoři nevyzývají k opuštění lékařské AI — doporučují měřit soukromí na úrovni jednotlivých pacientů, řídit přístup k modelům a používat diferenciální soukromí. Nepříjemné jádro výsledku: „soukromý v průměru“ není zárukou soukromí a selhává právě u pacientů, kteří už jsou chráněni nejméně.

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

Záruka soukromí je slib daný člověku, ne průměrnému člověku

Když se model lékařské AI nazývá „chránící soukromí“, toto tvrzení obvykle spočívá v jednom čísle: v průměru je malá šance určit, zda data konkrétního pacienta byla součástí tréninkové sady. To zní uklidňujícím dojmem. Nenápadný, ale nepříjemný závěr této práce zní: je to špatně zvolená metrika.

Soukromí není průměrné. Je to slib daný každému jednotlivci — že zařazení do tréninkové sady je později nevystaví nebezpečí. A průměr může tento slib splnit téměř pro každého, zatímco ho pro pár úplně poruší. Výzkumníci proto měřili riziko soukromí u jednotlivých pacientů, v dosud nepoužitém rozlišení, a zjistili přesně to: modely, které vypadají jako bezpečné, mohou téměř dokonale odhalovat členství konkrétních osob — a osobami, jejichž členství odhalují, jsou nepoměrně často lidé, kteří jsou již nejméně chráněni.

Co autoři udělali

Tým vedený Moritzem Knollem, s Danielem Rückertem (oba z Technické univerzity v Mnichově) a Georgem Kaissisem (Hasso Plattner Institute), spolu s kolegy z Imperial College London — vzal známou hrozbu pro soukromí, **útok na odvození členství**, a položil otázku jinak. Místo otázky “jak často je tento útok v průměru úspěšný v celé datové sadě?”, ptali se “jak ohrožen je *tento konkrétní pacient*?”

Tuto analýzu na úrovni jednotlivých pacientů provedli napříč **sedmi zavedenými soubory lékařských dat**, pokrývajícími velmi odlišné typy dat — lékařské zobrazování, elektrokardiogramy a elektronické zdravotní záznamy — a celkem na 200 modelech. U každého pacienta v každém modelu odhadli, jak jistě útočník dokáže poznat, zda záznam dané osoby byl součástí tréninkových dat, a poté výsledky rozdělili podle skupin: stav nemoci, rasa uváděná samotným pacientem, pojištění, pohlaví a zobrazovací protokol.

> **Co vlastně znamená útok na odvození členství**
>
> Únik informací je tu jemnější než „model vypíše váš záznam“. Jde o *členství* — samotný fakt, že vaše data byla ve trénovací sadě.
> Začněte tím, co jeden z těchto modelů obvykle dělá. Podáte mu vyšetření — například rentgen hrudníku — a on vám vrátí pravděpodobnosti: 78% pravděpodobnost zápalu plic, spolu s výsledky kardiomegalie, otoku, konsolidace. Ta každodenní diagnostická odpověď je *jediná* věc, kterou útok používá. Nic exotického.
> Slabina, na kterou se spoléhá, je tato: model je obvykle o něco **jistější** ohledně přesných příkladů, na kterých byl trénován, než ohledně těch, které nikdy neviděl. Představte si studenta, který tajně viděl zkoušku předem — u procvičených otázek přicházejí odpovědi trochu příliš rychle a až příliš sebejistě. Odvozovat z této stopy, zda konkrétní záznam patřil mezi tréninkové příklady, se nazývá „odvození členství“.

Takže tady je útok, krok za krokem. Někdo chce vědět, zda byl sken konkrétní osoby použit k trénování modelu. Nežádají model o záznam — už mají kandidátní snímek (snímek dané osoby nebo jeho blízkou kopii). Jednou jej odešlou, jako běžnou diagnostickou žádost, a poznamenají, jak je odpověď jistá. Pak zkontrolují, zda míra jistoty odpovídá spíše modelu, který na snímku *byl* trénován, nebo modelu, který na něm trénován *nebyl* — srovnání, které mohou levně provést tím, že natrénují vlastního zástupce ("referenční model"), aby zjistili, jak vypadá ta typická přehnaná jistota na běžném počítači bez zvláštního hardwaru. Pokud si je skutečný model u tohoto snímku neobvykle jistý — jako by ho rozpoznal — je to silný signál, že sken byl v trénovacích datech.

Znepokojující je, že nic na požadavku nevypadá jako útok: je to stejný diagnostický dotaz, jaký by provedl klinik, a signál soukromí je skrytý uvnitř běžné předpovědi. Právě proto je riziko reálné — ačkoli bylo zatím prokázáno podle laboratorních předpokladů, nikoli přímo v reálném provozu.

Proč je členství důležité, když samotný záznam nikdy neunikl? Protože *členství je fakt*. Pokud byl model trénován na pacientech, kteří podstoupili specifickou imunoterapii rakoviny, pak potvrzení, že v něm je váš záznam, odhaluje, že jste pravděpodobně měli tento typ rakoviny — což by pojišťovna nebo zaměstnavatel nikdy neměli odvodit. Obsah zůstává skrytý; uniká samotný fakt členství.

Autoři tyto předpoklady jasně stanovili — přístup k běžným predikcím modelu, kandidátní záznam a referenční model útočníka — nikoli jako návod, ale aby bylo možné hrozbu rozumně odhadnout, nikoli jen přehlížet.

Útok na odvození členství se může skrývat v běžné predikci

Útočník nežádá o záznam. Už má kandidáta a model se dotáže jednou.

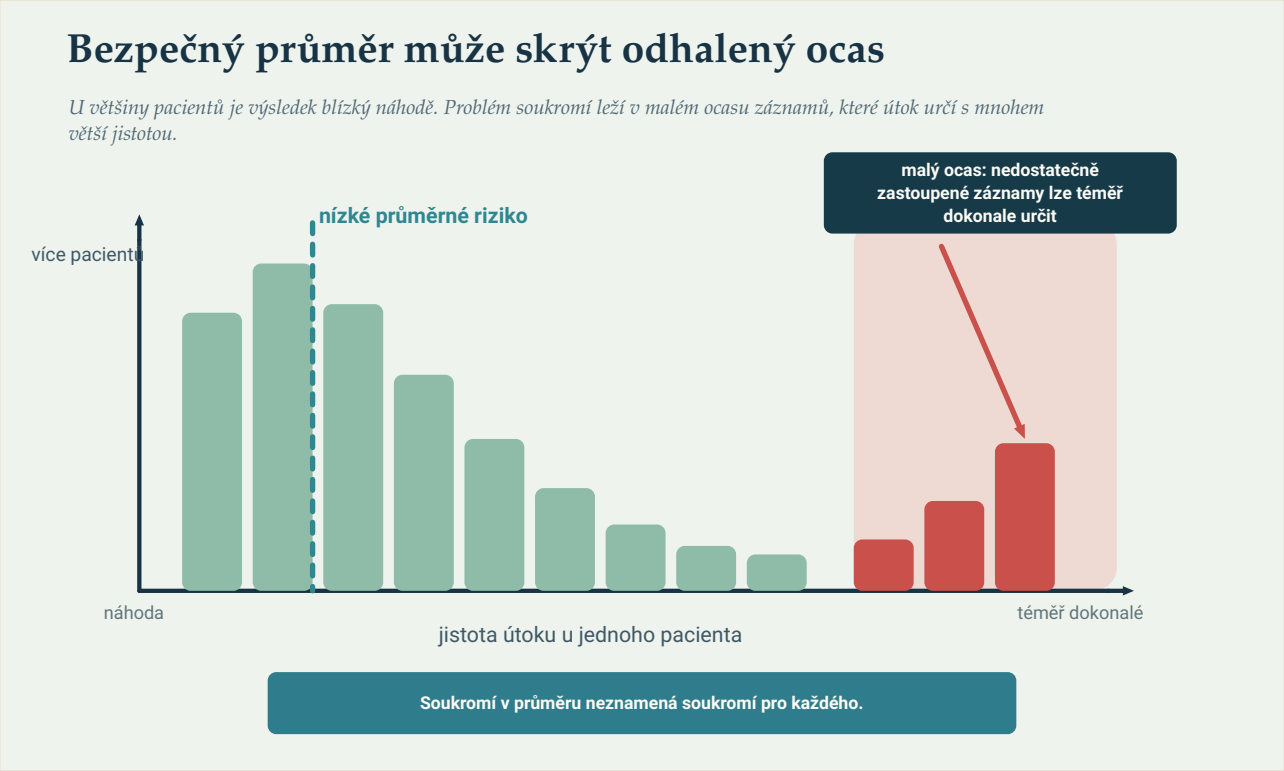
Pouze princip: žádný pseudokód, práh útoku ani návod.

Útok nepotřebuje speciální API pro ochranu soukromí. Běžný diagnostický dotaz může odhalit signál členství, pokud je model neobvykle jistý záznamem, který již viděl.

Co našli

Tři nálezy, každý ostřejší než ten předchozí.

Průměry skrývají ohrožené. Pokud se měří souhrnně — jak je to obvyklé — mnoho těchto modelů působí na první pohled bezpečně: u většiny pacientů si útok nevede lépe než náhoda. Pohled na úrovni jednotlivých pacientů ukázal jiný příběh. Jak Knolle uvedl v oznámení studie, předchozí hodnocení "vždy měřila pouze průměrné riziko u všech pacientů. Poprvé jsme zkoumali riziko na úrovni jednotlivých pacientů — a obraz je zcela odlišný." U některých jedinců útok uspěje **téměř dokonale** — autoři jej měří jako AUC útoku **0,95 nebo více** (0,5 je hod mincí, 1,0 je dokonalý výsledek) — i když průměr v celé datové sadě nevypadal lépe než náhoda.



Průměr může být blízko náhody, zatímco malý počet záznamů zůstává výrazně zranitelnější. Smyslem článku je, že soukromí musí být kontrolováno na úrovni pacienta, ne jen jako celek.

Expozice je nerovnoměrná — a dopadá na nedostatečně zastoupené skupiny. Pacienti nejzranitelnější vůči téměř dokonalému útoku byli systematicky ti ve skupinách **v datech nedostatečně zastoupeni**: příslušníci etnických menšin, fenotypy vzácných onemocnění, neobvyklé zobrazovací charakteristiky. V souboru elektronických zdravotních záznamů se černošští pacienti objevovali **o 31 % častěji** než se očekávalo mezi nejzranitelnějšími záznamy; v souboru mamografických snímků byly snímky označené jako podezřelé na malignitu v této vysoce rizikové skupině nadměrně zastoupeny

výrazným **+1 179 %**. (Tyto dvě údaje jsou příklady, nikoli celkový obraz — práce uvádí podobnou nerovnost napříč několika skupinami — a vyjadřují *relativní* nadměrné zastoupení v malém ocasu rozdělení s extrémním rizikem: o kolik častěji se tito pacienti objevují mezi nejvíce vystavenými záznamy, nikoli podíl černých nebo podezřelých pacientů, kteří jsou vystaveni.) Model má méně podobných příkladů, mezi nimiž by se tito pacienti mohli ztratit, takže jejich záznamy vynikají — a právě toto vyčnívání útok na odvození členství zachycuje. Selhání soukromí není náhodné; soustředí se na lidi, kteří už jsou na okraji.

Větší modely to zhoršují. Počet pacientů vystavených téměř dokonalým útokům prudce vzrostl s velikostí modelu — v jednom dermatologickém datovém souboru vzrostl podíl z prakticky nuly v nejmenším modelu na přibližně **jednoho z deseti pacientů** v největším. Jak se modely lékařské AI zvětšují a zvyšují schopnosti — směr, kterým se celý obor ubírá — autoři varují, že toto konkrétní riziko roste, nikoli snižuje.

Rückertovo shrnutí je přímočaré: “Toto není tolerovatelné riziko. Zdravotní data jsou velmi citlivá.”

Proč je “průměrně bezpečný” špatný test

Je lákavé číst nízké průměrné riziko jako potvrzení, že je vše v pořádku. Hlavní lekcí článku je, že průměrování zde není jen nepřesné — znamená to měření špatné věci.

Záruka soukromí má smysl pouze tehdy, pokud platí pro osobu nejvíce ohroženou, nikoli pro osobu uprostřed. Model, kde je 999 z 1 000 pacientů neidentifikovatelných, ale jeden lze téměř dokonale identifikovat, není “99,9 % soukromý” v žádném smyslu, který by byl pro tuto osobu podstatný — a pokud je tím člověkem předvídatelně pacient s vzácnou nemocí nebo pacient z etnické menšiny, metrika není jen neúplná, ale tiše diskriminační. Hlásí bezpečnost pro většinu a nazývá ji bezpečnost pro všechny.

Práce proto mění otázku: od *jak soukromý je tento model v průměru?* na *který pacient je nejvíce ohrožen a kdo to je?* To jsou různé otázky a jen druhá je otázka soukromí.

Co to *ne*dokazuje — ani netvrdí

- Neříká, že by měla být lékařská AI opuštěna. Autoři to chápou jako zmírnění, nikoli ústup: změřte a opravte riziko, nepřestávejte budovat.
- To **neznamená**, že každý lékařský model uniká, nebo že nějaký nasazený model byl napaden. Ukazuje to, že *riziko existuje a je nerovnoměrně rozdělené*, a že standardní agregované metriky ho přehlížejí.
- To neznamená, že vaše záznamy jsou už zveřejněné. Útok vyžaduje specifické podmínky — přístup k modelu, kandidátní záznam a vlastní infrastrukturu útočníka — nikoli náhodnou schopnost.
- Neukazuje únik *obsahu* pacientských dat. Co uniká, je členství — fakt účasti — což je nebezpečné z jiného důvodu, ne proto, že by se celý záznam přímo odhalil.

- Nesnižuje rozdíly na jedno hlavní číslo. Silným, reprodukovatelným výsledkem je *vzor* — agregované metriky podhodnocují individuální riziko a nedostatečně zastoupení pacienti ho nesou nejvíce — napříč datovými sadami a stovkami modelů.

Jak silné jsou důkazy?

Jedná se o empirický, metodologický výsledek a velmi robustní. Vzorec — agregované metriky soukromí systematicky podhodnocují riziko na pacienta, zbytkové riziko se soustředí na nedostatečně zastoupené skupiny a zhoršuje se s velikostí modelu — se drží napříč **sedmi datovými sadami různých typů dat** a velkým počtem modelů na datovou sadu. Právě tato šíře činí tvrzení o měření věrohodným, nikoli artefaktem z jedné datové sady.

Dvě poctivé výhrady. Za prvé, jde o ukázku *rizika*, měřené spouštěním útoků, které autoři sami vytvořili; charakterizuje, jak dobře by schopný útočník *mohl* uspět za předpokladů, nikoli jak často dochází ke skutečným útokům. Za druhé, živá čísla — téměř dokonalá úspěšnost útoků u některých pacientů — popisují nejhůře postižené jedince *záměrně*; právě v tom spočívá smysl analýzy a měly by se číst jako „chvost rozdělení je mnohem těžší, než naznačuje průměr“, ne jako “většina pacientů je vystavena.”

Správný postoj není ani alarm, ani odmítnutí: pečlivá, vícedatová studie ukazuje, že standardní způsob certifikace soukromí v lékařské AI je slepý vůči vlastním nejhorším případům — a že tyto nejhorší případy dopadají na pacienty s nejmenší rezervou.

Proč je to důležité

Dvě věci z toho dělají víc než jen technickou poznámku pod čarou.

Za prvé, mění to, co by měl pojem „ochrana soukromí“ smět znamenat. Pokud je model vydán s průměrným skóre ochrany soukromí, toto skóre může být skutečně nízké a přesto skrývat podskupinu pacientů, kteří jsou téměř dokonale rozpoznatelní útokem na členství. Praktický požadavek práce je konkrétní: před zveřejněním posoudit riziko soukromí **na pacienta**, kontrolovat, kdo může přistupovat k nasazeným modelům, a použít techniky jako **diferenciální soukromí** — malý, pečlivě kalibrovaný šum přidaný během tréninku, který otupuje útoky na členství, za skutečnou a řízenou cenu v podobě nižší užitečnosti modelu (kompromis mezi soukromím a užitečností je explicitní, nikoli zdarma). „Zkontrolovali jsme průměr“ by se mělo přestat počítat jako kontrola.

Za druhé, propojuje soukromí se spravedlností. Stejné skupiny, které jsou v lékařských datech nedostatečně zastoupeny — a tedy už jsou zdravotně nedostatečně obsluhovány — se ukážou být těmi, jejichž soukromí AI chrání nejméně. Obor, který tvrdě pracuje na první nerovnosti, nemůže druhou považovat za oddělení někoho jiného. Jsou to ti samí lidé.

Uklidňující příběh o soukromí v oblasti lékařské AI — *měřili jsme to, průměr je nízký, jsme v pořádku* — je příběh, který tato práce rozebírá. Ne proto, aby někoho odradil od technologie, ale aby se standard

posunul tam, kam patří: slib, který můžete dodržet jen tehdy, pokud jste ho zkontrolovali u osoby, která by mohla být nejvíce zraněna.

Stručné shrnutí

Útoky na základě odvození členství se snaží zjistit, zda záznam konkrétní osoby byl v tréninkových datech AI modelu — a protože členství samo o sobě může být odhalující (například že jste měli určitou nemoc), jedná se o skutečný únik soukromí, i když se základní záznam nikdy neobjeví. Předchozí práce měřily, jak často takové útoky uspějí *v průměru* v datové sadě. Tato studie měřila toto riziko *na pacienta*, napříč sedmi soubory lékařských dat (zobrazování, EKG, elektronické záznamy) a v mnoha modelech pro každou z nich, a zjistila, že agregované metriky riziko výrazně podhodnocují: některé jedince lze identifikovat téměř dokonale (AUC útoku $\geq 0{,}95$), i když průměr celého souboru dat vypadá bezpečně; nejzranitelnější jsou systematicky lidé z nedostatečně zastoupených skupin (příslušníci etnických menšin, vzácná onemocnění, neobvyklé zobrazovací metody); problém navíc roste s velikostí modelu. Autoři neargumentují pro opuštění lékařské AI — prosazují měření soukromí na individuální úrovni, kontrolu přístupu k modelům a využívání diferenciálního soukromí. Závěr: "soukromé v průměru" není zárukou soukromí a lidé, u nichž selhává, jsou obvykle ti, kteří už jsou nejméně chráněni.

Kontrola bez příkras

Co práce ukazuje: Napříč sedmi soubory lékařských dat a mnoha modely je riziko odvození členství na pacienta u některých jedinců mnohem vyšší, než naznačují souhrnné metriky — až do téměř dokonalého úspěchu útoku (AUC $\geq 0{,}95$) — a toto zbytkové riziko nepoměrně dopadá na nedostatečně zastoupené skupiny a roste s velikostí modelu.

Co je pravděpodobné, ale není prokázáno: Že skuteční útočníci to už využívají proti nasazeným klinickým modelům; studie ukazuje *schopnost a její rozložení* za uvedených předpokladů, nikoli frekvenci útoků v reálném provozu.

Co neukazuje: Že by měla být opuštěna lékařská AI; že každý model uniká nebo že byl narušen nějaký konkrétní nasazený model; že je zveřejněn *obsah* záznamu pacienta (místo členství); jedinou číselnou míru nerovnosti — robustní výsledek je vzor, nikoli jedno číslo.

Hlavní omezení: Měří riziko nejhoršího případu prostřednictvím vlastních útoků autorů, nikoli pozorovaných incidentů; dramatická čísla popisují nejvíce exponované jedince podle záměru; a přesné rozdíly mezi jednotlivými skupinami jsou zobrazeny jako konzistentní vzorec napříč datovými sadami, nikoli redukovány na jedno číslo.

Kolik důvěry by měl mít běžný čtenář? Vysokou v to, že agregované metriky soukromí podhodnocují individuální riziko, že zbytkové riziko se soustředí na nedostatečně zastoupené skupiny pacientů a že se to zhoršuje s velikostí modelu — což je patrné v mnoha datových sadách a modelech.

Střední ohledně skutečné četnosti takových útoků, kterou tato studie neměří. Bezpečné čtení: nikoli „lékařská AI způsobuje únik vašich dat“ ani „soukromí je vyřešeno“, ale „standardní kontrola soukromí nevidí své nejhorší případy a ty dopadají na nejzranitelnější pacienty — kontrola se tedy musí změnit“.

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