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

# Medyczna AI może być prywatna średnio, a mimo to narażać konkretnych pacjentów — zwłaszcza słabo reprezentowanych

*Określenie modelu medycznej AI jako „chroniącego prywatność" zwykle opiera się na jednej wartości średniej. Autorzy tego badania twierdzą, że to niewłaściwy test. Badając ataki wnioskowania o przynależności — ujawniające, czy rekord konkretnej osoby znalazł się w danych treningowych modelu, a tym samym mogące zdradzić, że miała ona określoną chorobę — mierzyli ryzyko osobno dla każdego pacjenta, nie tylko zbiorczo. Analiza objęła siedem medycznych zbiorów danych (obrazy, EKG i dokumentację elektroniczną) oraz wiele modeli dla każdego z nich. Wzorzec był następujący: modele wyglądające średnio na bezpieczne mogą nadal pozwalać atakującemu niemal bezbłędnie rozpoznawać konkretne osoby (AUC ataku $\geq 0{,}95$); narażeni systematycznie należą do grup słabo reprezentowanych (mniejszości etniczne, rzadkie choroby, nietypowe obrazy); problem narasta też wraz z wielkością modeli. Autorzy nie postulują rezygnacji z medycznej AI — zalecają mierzenie prywatności na poziomie pacjenta, kontrolę dostępu do modeli i stosowanie prywatności różnicowej. Niewygodne sedno: „prywatny średnio" nie jest gwarancją prywatności i zawodzi właśnie pacjentów już najmniej chronionych.*

---

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

# Gwarancja prywatności jest obietnicą złożoną osobie, nie średniej

Gdy model medycznej AI określa się jako „chroniący prywatność", twierdzenie zwykle opiera się na jednej liczbie: wśród wszystkich pacjentów, których dane posłużyły do treningu, średnia szansa odgadnięcia udziału konkretnej osoby jest niska. Brzmi to uspokajająco. Mniej widoczny, ale niewygodny wniosek tej pracy jest taki, że to niewłaściwa liczba.

Prywatność nie jest średnią. Jest obietnicą składaną każdej osobie — że obecność w zbiorze treningowym nie wróci, by ją narazić. Średnia może dotrzymywać tej obietnicy niemal wszystkim, a całkowicie złamać ją wobec kilku osób. Badacze zmierzyli ryzyko prywatności osobno dla każdego pacjenta, z rozdzielczością wcześniej niestosowaną, i znaleźli właśnie to: modele wyglądające bezpiecznie w ujęciu zbiorczym mogą niemal bezbłędnie ujawniać udział konkretnych osób — a ujawniają nieproporcjonalnie często tych, którzy już wcześniej byli najsłabiej chronieni.

# Co zrobili autorzy

Zespół kierowany przez Moritza Knollego, z Danielem Rückertem (obaj z Uniwersytetu Technicznego w Monachium), Georgiem Kaissisem (Hasso Plattner Institute) i współpracownikami z Imperial College London, wziął znane zagrożenie dla prywatności zwane **atakiem wnioskowania o przynależności** i zmienił pytanie. Zamiast pytać „jak często atak średnio udaje się w całym zbiorze?", badacze zapytali: „jak bardzo narażony jest *ten konkretny pacjent*?".

Przeprowadzili analizę na poziomie pacjenta w **siedmiu uznanych medycznych zbiorach danych**, obejmujących bardzo różne typy informacji — obrazowanie medyczne, elektrokardiogramy i elektroniczną dokumentację medyczną — oraz po 200 modeli dla każdego z nich. Dla każdego pacjenta w każdym modelu oszacowali, z jaką pewnością atakujący może ustalić, czy jego dokumentacja była częścią danych treningowych, a następnie rozłożyli wyniki według grup: stanu chorobowego, deklarowanej rasy, ubezpieczenia, płci i protokołu obrazowania.

> **Czym naprawdę jest atak wnioskowania o przynależności**
>
> Wyciek jest subtelniejszy niż „model wypisuje twoją dokumentację". Dotyczy *przynależności* — samego faktu, że dane znalazły się w zbiorze treningowym.
> Zacznijmy od zwykłego działania takiego modelu. Przekazujemy mu obraz, na przykład zdjęcie rentgenowskie klatki piersiowej, a on zwraca prawdopodobieństwa: 78% szans na zapalenie płuc oraz oceny kardiomegalii, obrzęku i konsolidacji. Atak wykorzystuje *wyłącznie* tę codzienną odpowiedź diagnostyczną. Nic niezwykłego.
> Słabość polega na tym, że model jest zwykle nieco **bardziej pewny** dokładnie tych przykładów, na których go trenowano, niż przykładów nigdy wcześniej niewidzianych. Przypomina ucznia, który potajemnie zobaczył egzamin: przy przećwiczonych pytaniach odpowiedzi przychodzą odrobinę za szybko i zbyt pewnie. „Wnioskowanie o przynależności" oznacza ustalenie na

podstawie tej wskazówki, czy konkretny zapis należał do przykładów badanych przez model. Atak przebiega więc krok po kroku. Ktoś chce wiedzieć, czy obraz konkretnej osoby wykorzystano do treningu modelu. **Nie** prosi modelu o sam zapis — ma już kandydacki obraz (należący do tej osoby albo jego bliską kopię). Wysyła go raz jako zwykłe żądanie diagnostyczne i zapisuje pewność odpowiedzi. Następnie sprawdza, czy przypomina ona bardziej model, który *trenował* na tym obrazie, czy taki, który go *nie widział*. Porównanie może tanio przygotować, trenując własny substytut („model referencyjny"), aby nauczyć się charakterystycznego wzorca nadmiernej pewności, na zwykłym komputerze bez specjalnego sprzętu. Jeśli rzeczywisty model jest wyjątkowo pewny tego obrazu — jakby go rozpoznawał — jest to mocny dowód, że obraz należał do danych treningowych.

Niepokojące jest to, że żaden element żądania nie wygląda jak atak: jest to to samo zapytanie diagnostyczne, które zadałby lekarz, a sygnał prywatności kryje się wewnątrz zwykłej predykcji. Właśnie dlatego ryzyko jest konkretne — choć dotychczas wykazano je przy określonych założeniach laboratoryjnych, a nie przyłapano w praktyce.

Dlaczego przynależność ma znaczenie, jeśli sam zapis nigdy nie wycieka? Ponieważ *przynależność jest faktem*. Jeżeli model trenowano na pacjentach otrzymujących określoną immunoterapię nowotworową, potwierdzenie obecności danego zapisu ujawnia, że osoba prawdopodobnie miała ten nowotwór — informacji takiej ubezpieczyciel ani pracodawca nie powinien nigdy móc wywnioskować. Treść pozostaje zamknięta; wydostaje się fakt przynależności.

Autorzy jasno przedstawiają założenia — dostęp do zwykłych predykcji modelu, kandydacki zapis i własny model referencyjny atakującego — nie jako instrukcję, lecz po to, by zagrożenie można było rzeczowo przeanalizować zamiast je zbywać.

## Atak wnioskowania o przynależności może kryć się w zwykłej predykcji

*Atakujący nie widzi dokumentacji. Ma już rekord kandydata i tylko raz odpytuje model.*


*Tylko mechanizm: brak pseudokodu, brak progu ataku, brak przepisu.*

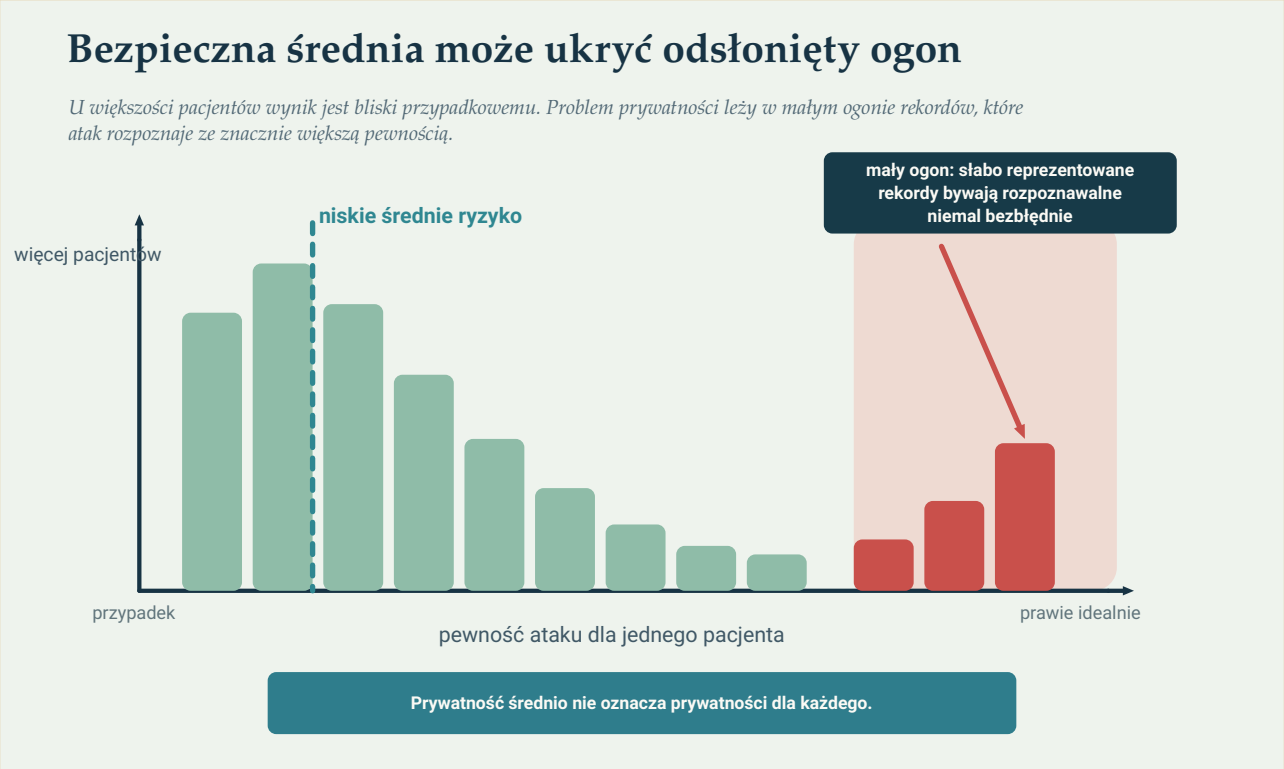
Atak nie potrzebuje specjalnego interfejsu prywatności. Zwykłe zapytanie diagnostyczne może ujawnić sygnał przynależności, jeśli model jest niezwykle pewny wobec zapisu, który wcześniej widział.



# Co odkryli

Trzy wyniki, każdy ostrzejszy od poprzedniego.

**Średnie ukrywają narażonych.** Przy pomiarze zbiorczym — stosowanym zazwyczaj — wiele modeli wygląda uspokajająco bezpiecznie: dla większości pacjentów atak nie wypada lepiej niż losowanie. Widok na poziomie pacjenta opowiada inną historię. Jak ujął to Knolle w komunikacie o badaniu, wcześniejsze oceny „zawsze mierzyły tylko średnie ryzyko dla wszystkich pacjentów. Po raz pierwszy zbadaliśmy ryzyko na poziomie pojedynczych pacjentów — i obraz jest zupełnie inny". Dla niektórych osób atak udaje się **niemal bezbłędnie** — autorzy mierzą to jako AUC ataku równe **0,95 lub więcej** (0,5 odpowiada rzutowi monetą, 1,0 oznacza wynik bezbłędny) — nawet gdy średnia całego zbioru nie przewyższa losowania.




Średnia może znajdować się blisko poziomu losowego, podczas gdy niewielki ogon zapisów pozostaje znacznie bardziej narażony. Praca pokazuje, że prywatność należy sprawdzać na poziomie pacjenta, nie tylko zbiorczo.



**Narażenie jest nierówne — i dotyka grup niedostatecznie reprezentowanych.** Pacjenci najbardziej podatni na niemal bezbłędny atak systematycznie należeli do grup **niedostatecznie reprezen-**

**towanych w danych**: mniejszości etnicznych, osób z rzadkimi fenotypami chorób i nietypowymi cechami obrazowania. W zbiorze elektronicznej dokumentacji czarni pacjenci pojawiali się wśród najbardziej narażonych zapisów **o 31% częściej**, niż oczekiwano; w zbiorze mammograficznym obrazy oznaczone jako podejrzane o złośliwość były w tej strefie zagrożenia nadreprezentowane aż o **+1179%**. (Te dwie wartości są przykładami, nie całym obrazem — praca opisuje podobne odchylenie w kilku grupach — i oznaczają *względną* nadreprezentację w niewielkim ogonie skrajnego ryzyka: jak często tacy pacjenci pojawiają się wśród najbardziej narażonych zapisów, a nie jaki odsetek czarnych pacjentów albo pacjentów z podejrzaną mammografią jest narażony.) Model ma mniej podobnych przykładów, wśród których mógłby ukryć takich pacjentów, dlatego ich dokumentacja się wyróżnia — a właśnie wyróżnianie wykrywa atak wnioskowania o przynależności. Naruszenie prywatności nie jest losowe; skupia się na osobach już znajdujących się na marginesie.

**Większe modele pogarszają problem.** Liczba pacjentów narażonych na niemal bezbłędne ataki rosła gwałtownie wraz z pojemnością modelu — w jednym zbiorze dermatologicznym udział zwiększył się od praktycznie zera w najmniejszym modelu do około **jednej osoby na dziesięć** w największym. Autorzy ostrzegają, że wraz ze wzrostem rozmiaru i możliwości medycznych modeli AI — czyli kierunkiem rozwoju całej dziedziny — to konkretne ryzyko staje się poważniejsze, a nie mniejsze.

Podsumowanie Rückerta jest dosadne: „To ryzyko nie jest do przyjęcia. Dane zdrowotne są wyjątkowo wrażliwe".

## Dlaczego „bezpieczny średnio" to niewłaściwy test

Kusi, by odczytać niskie średnie ryzyko jako świadectwo pełnego zdrowia. Główna lekcja pracy mówi, że uśrednianie nie jest tutaj jedynie niedokładne — mierzy niewłaściwą rzecz.

Gwarancja prywatności ma sens tylko wtedy, gdy obowiązuje osobę najbardziej zagrożoną, a nie środkową. Model, w którym 999 z 1000 pacjentów nie można rozpoznać, lecz jednego można zidentyfikować niemal bezbłędnie, nie jest „prywatny w 99,9%" w żadnym sensie istotnym dla tej jednej osoby. Jeśli tą osobą jest w przewidywalny sposób pacjent z rzadką chorobą albo należący do mniejszości etnicznej, wskaźnik jest nie tylko niepełny, lecz po cichu dyskryminujący. Informuje o bezpieczeństwie większości i nazywa je bezpieczeństwem wszystkich.

Praca wymusza zmianę pytania z *jak prywatny jest ten model średnio?* na *który pacjent jest najbardziej narażony i kim on jest?* To różne pytania, a tylko drugie naprawdę dotyczy prywatności.

## Czego to *nie dowodzi* — ani nie twierdzi

- Praca **nie mówi**, że należy porzucić medyczną AI. Autorzy proponują ograniczenie ryzyka, nie odwrót: mierzyć je i usuwać, a nie przestać budować.
- **Nie oznacza**, że każdy model medyczny ujawnia dane ani że którykolwiek wdrożony model został zaatakowany. Pokazuje, że *ryzyko istnieje i jest nierówno rozłożone*, a standardowe wskaźniki zbior-

cze go nie zauważają.

- **Nie oznacza**, że dokumentacja konkretnego czytelnika jest już narażona. Atak wymaga określonych warunków — dostępu do modelu, kandydackiego zapisu i własnej infrastruktury atakującego — nie jest przypadkową możliwością.
- **Nie pokazuje** wycieku *treści* danych pacjenta. Wyciekowi podlega przynależność — fakt udziału — groźny z innego powodu, a nie dlatego, że model wyrzuca całą dokumentację.
- **Nie sprowadza** nierówności do jednej liczby z nagłówka. Mocnym, powtarzalnym wynikiem jest *wzorzec* obecny w zbiorach i setkach modeli: wskaźniki zbiorcze zaniżają ryzyko indywidualne, a największy ciężar ponoszą pacjenci niedostatecznie reprezentowani.

## Jak mocne są dowody?

Jest to empiryczny wynik metodologiczny i jest solidny. Wzorzec — systematyczne zaniżanie ryzyka na poziomie pacjenta przez zbiorcze wskaźniki prywatności, koncentracja pozostałego ryzyka w grupach niedostatecznie reprezentowanych i jego wzrost wraz z pojemnością modelu — utrzymywał się w **siedmiu zbiorach różnych typów danych** i dużej liczbie modeli na zbiór. Właśnie ten zakres czyni twierdzenie o pomiarze wiarygodnym, a nie artefaktem jednego zestawu.

Dwa uczciwe zastrzeżenia. Po pierwsze, jest to demonstracja *ryzyka*, mierzona przez przeprowadzenie ataków zbudowanych przez samych autorów; opisuje, co zdolny atakujący *mógłby* osiągnąć przy określonych założeniach, a nie jak często rzeczywiste ataki się zdarzają. Po drugie, wyraziste liczby — niemal bezbłędny atak na niektórych pacjentów — z założenia opisują osoby w najgorszej sytuacji. Na tym polega sedno i należy je czytać jako „ogon jest znacznie cięższy, niż sugeruje średnia", a nie „większość pacjentów jest narażona".

Właściwym podejściem nie jest ani alarmizm, ani lekceważenie: to staranne badanie wielu zbiorów pokazujące, że standardowy sposób certyfikowania prywatności medycznej AI nie widzi własnych najgorszych przypadków — a przypadki te dotyczą pacjentów mających najmniej przestrzeni na kolejne ryzyko.

## Dlaczego to ma znaczenie

Dwie rzeczy czynią wynik czymś więcej niż przypisem technicznym.

Po pierwsze, zmienia on dopuszczalne znaczenie określenia „chroniący prywatność". Model może zostać udostępniony z rzeczywiście niskim średnim wskaźnikiem prywatności, a mimo to ukrywać podzbiór pacjentów niemal doskonale rozpoznawalnych przez atak wnioskowania o przynależności. Praktyczne żądanie pracy jest konkretne: przed udostępnieniem oceniać ryzyko prywatności **dla każdego pacjenta**, kontrolować dostęp do wdrożonych modeli i używać technik takich jak **prywatność różnicowa** — niewielki, starannie skalibrowany szum dodawany podczas treningu, który osłabia ataki na przynależność kosztem rzeczywistego, kontrolowanego spadku użyteczności modelu (kompromis między prywatnością a użytecznością jest jawny, nie darmowy). „Sprawdziliśmy średnią" nie

powinno już uchodzić za sprawdzenie.

Po drugie, praca splata prywatność ze sprawiedliwością. Te same grupy, które są niedostatecznie reprezentowane w danych medycznych — i dlatego już dziś gorzej obsługiwane przez medyczną AI — okazują się tymi, których prywatność AI chroni najsłabiej. Dziedzina usilnie pracująca nad pierwszą nierównością nie może traktować drugiej jako cudzej odpowiedzialności. Dotyczą tych samych ludzi.

Uspokajająca opowieść o prywatności medycznej AI — *zmierzyliśmy ją, średnia jest niska, wszystko w porządku* — zostaje w tej pracy rozebrana. Nie po to, by kogokolwiek odstraszyć od technologii, lecz by przesunąć standard tam, gdzie powinien być: obietnicy, której dotrzymanie można deklarować dopiero po sprawdzeniu jej wobec osoby najbardziej narażonej na szkodę.

# Zwięzłe podsumowanie

Ataki wnioskowania o przynależności próbują ustalić, czy dokumentacja konkretnej osoby znajdowała się w danych treningowych modelu AI — a ponieważ sama przynależność może coś ujawniać (na przykład przebycie określonej choroby), jest to rzeczywisty wyciek prywatności nawet wtedy, gdy bazowy zapis nigdy się nie pojawia. Wcześniejsze prace mierzyły, jak często ataki takie udają się *średnio* w zbiorze. To badanie zmierzyło wynik *dla każdego pacjenta* w siedmiu medycznych zbiorach danych (obrazowanie, EKG, elektroniczna dokumentacja) i wielu modelach dla każdego z nich. Wykazało, że wskaźniki zbiorcze poważnie zaniżają ryzyko: niektóre osoby można rozpoznać niemal bezbłędnie (AUC ataku $\geq 0{,}95$), nawet gdy średnia całego zbioru wygląda bezpiecznie; najbardziej narażeni systematycznie należą do grup niedostatecznie reprezentowanych (mniejszości etniczne, rzadkie choroby, nietypowe obrazowanie); a problem rośnie wraz z rozmiarem modelu. Autorzy nie postulują porzucenia medycznej AI — proponują pomiar prywatności na poziomie osoby, kontrolowanie dostępu do modelu i stosowanie prywatności różnicowej. Wniosek: „prywatny średnio" nie jest gwarancją prywatności, a osoby, wobec których zawodzi, zazwyczaj są już najsłabiej chronione.

# Kontrola bez upiększeń

**Co pokazuje praca:** W siedmiu medycznych zbiorach i wielu modelach dla każdego z nich ryzyko ataku wnioskowania o przynależności na poziomie pacjenta jest dla niektórych osób znacznie wyższe, niż sugerują wskaźniki zbiorcze — aż do niemal bezbłędnego ataku (AUC $\geq 0{,}95$) — a pozostałe ryzyko nieproporcjonalnie obciąża grupy niedostatecznie reprezentowane i rośnie wraz z pojemnością modelu.

**Co jest prawdopodobne, ale nieudowodnione:** Że rzeczywiści atakujący już wykorzystują tę metodę przeciw wdrożonym modelom klinicznym; badanie demonstruje *możliwość i jej rozkład* przy określonych założeniach, nie częstość ataków w praktyce.

**Czego praca nie pokazuje:** Że medyczną AI należy porzucić; że każdy model ujawnia informacje

albo konkretny wdrożony model został naruszony; że ujawniono *treść* dokumentacji pacjenta, a nie przynależność; jednego ilościowego wskaźnika nierówności — solidnym wynikiem jest wzorzec, nie jedna liczba.

**Główne ograniczenia:** Praca mierzy ryzyko najgorszego przypadku za pomocą ataków autorów, nie zaobserwowanych incydentów; dramatyczne liczby z założenia opisują najbardziej narażone osoby; dokładne nierówności między grupami przedstawiono jako spójny wzorzec w różnych zbiorach, a nie jedną wartość.

**Jak duże zaufanie powinien mieć czytelnik niespecjalista?** Wysokie do wniosku, że zbiorcze wskaźniki prywatności zaniżają ryzyko indywidualne, pozostałe ryzyko koncentruje się na niedostatecznie reprezentowanych pacjentach i rośnie z rozmiarem modelu — wykazano to w wielu zbiorach i modelach. Umiarkowane wobec rzeczywistej częstości takich ataków, której badanie nie mierzy. Bezpieczny odczyt brzmi nie „medyczna AI ujawnia twoje dane" ani „prywatność jest rozwiązana", lecz „standardowa kontrola prywatności nie widzi własnych najgorszych przypadków, a dotyczą one najbardziej narażonych pacjentów — więc kontrola musi się zmienić".

---

## Źródła

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