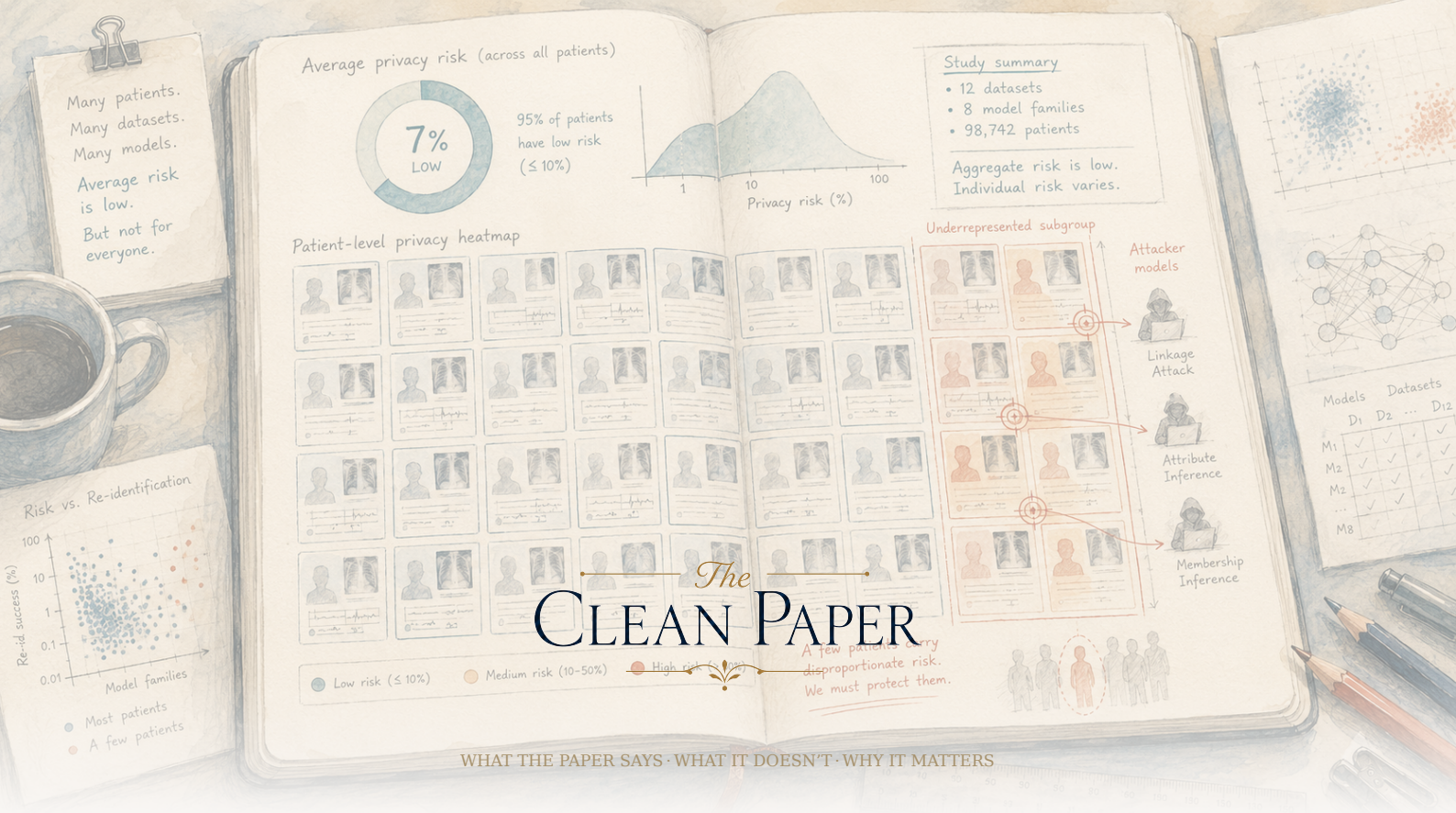


სამედიცინო AI შეიძლება საშუალოდ კონფიდენციალური იყოს და მაინც ამხელდეს კონკრეტულ პაციენტებს — განსაკუთრებით ნაკლებად წარმოდგენილ ჯგუფებს

„კონფიდენციალურობის დამცავი“ სამედიცინო AI-მოდელი ჩვეულებრივ ერთ საშუალო მაჩვენებელზე ფასდება. ეს კვლევა აჩვენებს, რომ ასეთი ტესტი საკმარისი არ არის. ავტორებმა membership inference attack-ის — შეტევის, რომელიც ადგენს, იყო თუ არა კონკრეტული ადამიანის ჩანაწერი მოდელის სასწავლო მონაცემებში — რისკი პაციენტის დონეზე გაზომეს შვიდ სამედიცინო მონაცემთა ნაკრებში (გამოსახულებები, ECG, ელექტრონული ჩანაწერები) და მრავალ მოდელში. შედეგი: საშუალოდ უსაფრთხოდ მოჩანებულ მოდელშიც ზოგი ადამიანი თითქმის სრულყოფილად იდენტიფიცირდება (attack AUC ≥ 0.95); ყველაზე დაუცველები სისტემატურად ნაკლებად წარმოდგენილი ჯგუფების წარმომადგენლები არიან; და რისკი მოდელის ზრდასთან ერთად მატულობს. ავტორები სამედიცინო AI-ის მიტოვებას არ ითხოვენ — ისინი მოითხოვენ კონფიდენციალურობის პაციენტის დონეზე შეფასებას, მოდელზე წვდომის კონტროლს და differential privacy-ის გამოყენებას. მთავარი აზრი: „საშუალოდ კონფიდენციალური“ ინდივიდუალური გარანტია არ არის.

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

კონფიდენციალურობის გარანტია დაპირებაა კონკრეტული ადამიანის მიმართ და არა საშუალო მაჩვენებლის მიმართ

როდესაც სამედიცინო AI-მოდელს „კონფიდენციალურობის დამცავს“ უწოდებენ, ეს განცხადება ჩვეულებრივ ერთ რიცხვს ეყრდნობა: ყველა იმ პაციენტზე საშუალოდ, რომელთა მონაცემებითაც მოდელი გაწვრთნეს, დაბალია ალბათობა, რომ ვინმემ დაადგინოს — შედიოდა თუ არა კონკრეტული ადამიანის ჩანაწერი სასწავლო მონაცემებში. ეს დამამშვიდებლად ჟღერს. ამ ნაშრომის მშვიდი, მაგრამ არასასიამოვნო დასკვნა ის არის, რომ სწორედ ესაა არასწორი მაჩვენებელი.

კონფიდენციალურობა საშუალო არ არის. ეს თითოეული ადამიანისადმი მიცემული დაპირებაა — რომ სასწავლო ნაკრებში ყოფნა მომავალში მის გასამჟღავნებლად არ შემოუბრუნდება. საშუალო მაჩვენებელს კი შეუძლია ეს დაპირება თითქმის ყველასთვის შეასრულოს და რამდენიმე ადამიანისთვის მთლიანად დაარღვიოს. მკვლევრებმა კონფიდენციალურობის რისკი პაციენტების მიხედვით, ცალ-ცალკე, აქამდე გამოუყენებელი სიზუსტით გაზომეს და სწორედ ეს აღმოაჩინეს: მოდელები, რომლებიც მთლიანობაში უსაფრთხოდ გამოიყურებიან, ზოგჯერ კონკრეტული ადამიანების სასწავლო ნაკრებში ყოფნას თითქმის უშეცდომოდ ამჟღავნებენ — და ეს ადამიანები არაპროპორციულად ხშირად სწორედ ის ჯგუფებია, რომლებიც ისედაც ყველაზე ნაკლებად არიან დაცული.

რა გააკეთეს ავტორებმა

ჯგუფმა — რომელსაც ხელმძღვანელობდა მორიც კნოლე, დანიელ რიუკერტთან (ორივე მიუნხენის ტექნიკური უნივერსიტეტიდან) და გეორგ კაისისთან (Hasso Plattner Institute) ერთად, ასევე Imperial College London-ის კოლეგების მონაწილეობით — აიღო კონფიდენციალურობის ცნობილი საფრთხე, რომელსაც **კუთვნილება დასკვნა attack**, ანუ სასწავლო ნაკრებში წევრობის დადგენის შეტევა ეწოდება, და კითხვა შეცვალა. ნაცვლად იმისა, რომ ეკითხათ: „საშუალოდ, რამდენად ხშირად მუშაობს ეს შეტევა მთელ მონაცემთა ნაკრებში?“, მათ იკითხეს: „კონკრეტულად *ეს პაციენტი* რამდენად დაუცველია?“

მათ პაციენტების მიხედვით ანალიზი ჩაატარეს **შვიდ დამკვიდრებულ სამედიცინო მონაცემთა ნაკრებზე**, რომლებიც ძალიან განსხვავებულ მონაცემებს მოიცავდა — სამედიცინო გამოსახულებებს, ელექტროკარდიოგრამებს და ელექტრონულ სამედიცინო ჩანაწერებს — და თითოეულ მათგანში 200 მოდელზე. თითოეული მოდელის თითოეული პაციენტისთვის მკვლევრებმა შეაფასეს, რამდენად თავდაჯერებით შეძლებდა თავდამსხმელი დაედგინა, იყო თუ არა ამ ადამიანის ჩანაწერი სასწავლო მონაცემებში, შემდეგ კი შედეგები ჯგუფების მიხედვით დაყვეს: დაავადების სტატუსი, თვითდეკლარირებული რასა, დაზღვევა, სქესი და გამოსახულების მიღების პროტოკოლი.

რას ნიშნავს სინამდვილეში კუთვნილება დასკვნა attack

აქ გაჟონვა უფრო დახვეწილია, ვიდრე „მოდელი თქვენს ჩანაწერს პირდაპირ გასცემს“. საუბარია მხოლოდ *წევრობაზე* — იმ ფაქტზე, რომ თქვენი მონაცემები სასწავლო ნაკრებში იყო.

დავიწყოთ იმით, რასაც ასეთი მოდელი ჩვეულებრივ აკეთებს. მაგალითად, აწვდით გულმკერდის რენტგენოგრამას და ის აბრუნებს ალბათობებს: პნევმონიის 78%-იან ალბათობას, ასევე შეფასებებს კარდიომეგალიაზე, შეშუპებაზე და კონსოლიდაციაზე. შეტევა მხოლოდ ამ ჩვეულებრივ დიაგნოსტიკურ პასუხს იყენებს. არაფერი განსაკუთრებული არ სჭირდება.

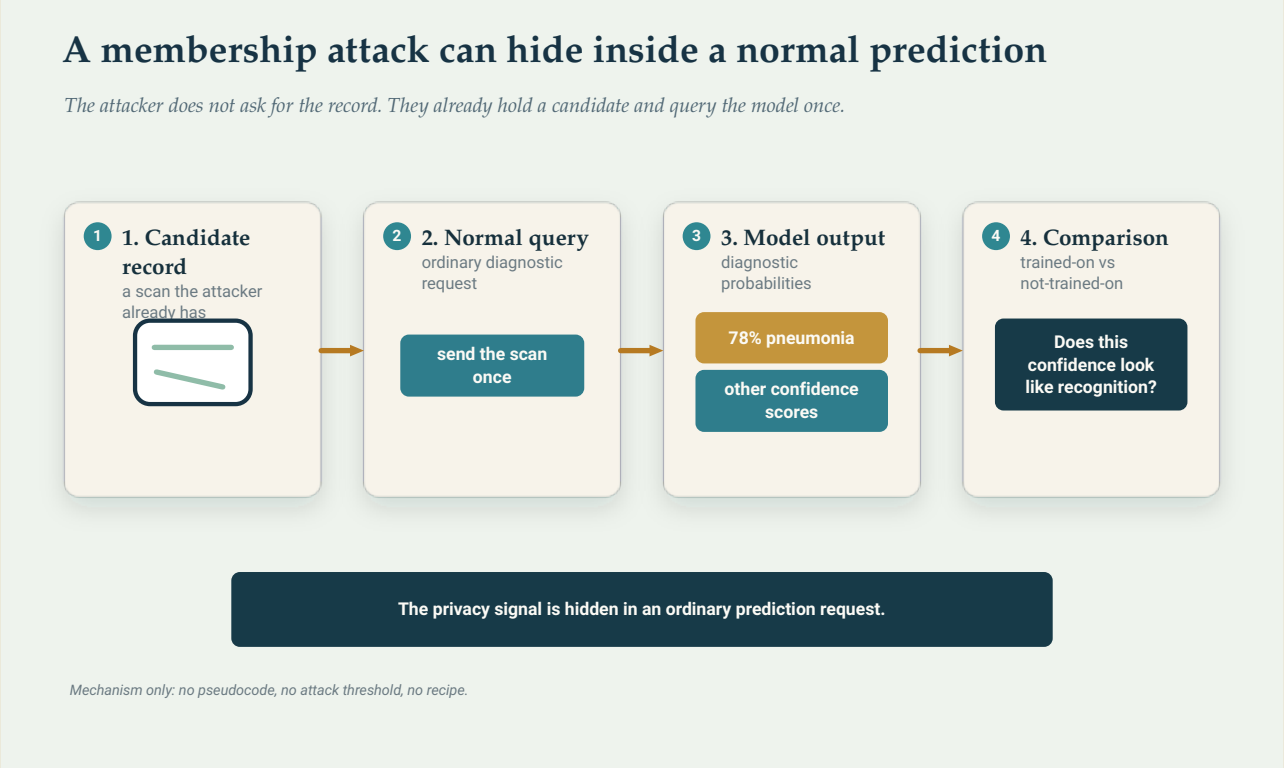
სისუსტე, რომელსაც ის ეყრდნობა, ასეთია: მოდელი, როგორც წესი, ოდნავ **უფრო თავდაჯერებულია** ზუსტად იმ მაგალითებზე, რომლებზეც გაწვრთნეს, ვიდრე იმ შემთხვევებზე, რომლებიც არასოდეს უნახავს. ეს ჰგავს სტუდენტს, რომელსაც გამოცდის კითხვები წინასწარ ჰქონდა ნანახი — ნაცნობ კითხვებზე პასუხები ცოტათი ზედმეტად სწრაფად და ზედმეტად თავდაჯერებით მოდის. ამ მინიშნებით იმის დადგენა, იყო თუ არა კონკრეტული ჩანაწერი მოდელის სასწავლო მაგალითებს შორის, სწორედ „კუთვნილება დასკვნა“-ს ნიშნავს.

შეტევა ეტაპობრივად ასე გამოიყურება. ვინმეს სურს გაიგოს, გამოიყენეს თუ არა კონკრეტული ადამიანის სკანი მოდელის გასაწვრთნელად. ის მოდელს ჩანაწერს **არ სთხოვს** — კანდიდატი სკანი უკვე აქვს, მაგალითად თავად ამ ადამიანის სკანი ან მისი ძალიან ახლო ასლი. მას მოდელში ერთხელ აგზავნის, როგორც ჩვეულებრივ დიაგნოსტიკურ მოთხოვნას, და აკვირდება პასუხის თავდაჯერებულობას. შემდეგ ამ თავდაჯერებულობას ადარებს იმას, როგორ იქცევა მოდელი სკანზე, რომელიც სწავლებისას *უნახავს*, და სკანზე, რომელიც *არ უნახავს*. ასეთი შედარების სწავლა თავდამსხმელს შედარებით იაფად შეუძლია საკუთარი დამხმარე „საკონტროლო მოდელი“-ის გაწვრთნით, ჩვეულებრივ კომპიუტერზე და სპეციალური აპარატურის გარეშე. თუ ნამდვილი მოდელი ამ კონკრეტულ სკანზე უჩვეულოდ დარწმუნებულია — თითქოს მას „ცნობს“ — ეს ძლიერი მინიშნებაა, რომ სკანი სასწავლო მონაცემებში იყო.

შემაშფოთებელი ისაა, რომ მოთხოვნა გარედან შეტევას საერთოდ არ ჰგავს: ის ზუსტად ისეთივე დიაგნოსტიკური შეკითხვაა, როგორსაც კლინიცისტი დასვამდა, ხოლო კონფიდენციალურობის სიგნალი ჩვეულებრივ პროგნოზშია დამალული. სწორედ ამიტომ არის რისკი კონკრეტული — მიუხედავად იმისა, რომ ჯერჯერობით ის აღწერილი ლაბორატორიული პირობების ფარგლებშია დემონსტრირებული და რეალურ გარემოში დაფიქსირებულ შეტევად არ გვინახავს.

რატომ აქვს მნიშვნელობა წევრობას, თუ თავად ჩანაწერი არასოდეს ჟონავს? იმიტომ, რომ *წევრობა თავისთავად ფაქტია*. თუ მოდელი გაწვრთნილი იყო იმ პაციენტებზე, რომლებმაც კონკრეტული კიბოს იმუნოთერაპია მიიღეს, იმის დადასტურება, რომ თქვენი ჩანაწერი ამ მოდელის სასწავლო მონაცემებში იყო, შეიძლება მიანიშნებდეს, რომ თქვენც სწორედ ეს კიბო გქონდათ — ინფორმაცია, რომლის დასკვნის გაკეთებაც დამსაქმებელს ან სადაზღვევო კომპანიას არ უნდა შეეძლოს. ჩანაწერის შინაარსი დახურული რჩება; გასაჯაროებულია კუთვნილების ფაქტი.

> ავტორები თავდასხმის პირობებს მკაფიოდ ჩამოთვლიან — მოდელის ჩვეულებრივ პროგნოზებზე წვდომა, კანდიდატი ჩანაწერი და თავდამსხმელის საკუთარი საკონტროლო მოდელი — არა იმისთვის, რომ ინსტრუქცია მისცენ, არამედ იმისთვის, რომ საფრთხის შეფასება კონკრეტულ პირობებს დაეყრდნოს და არა ბუნდოვან ვარაუდებს.



შეტევას სპეციალური კონფიდენციალურობის API არ სჭირდება. ჩვეულებრივმა დიაგნოსტიკურმა მოთხოვნამ შეიძლება გააჟონოს სასწავლო ნაკრებში წევრობის სიგნალი, თუ მოდელი უჩვეულოდ თავდაჯერებულია იმ ჩანაწერზე, რომელიც ადრე უნახავს.

რა აღმოაჩინეს

სამი შედეგი მიიღეს, და თითოეული წინაზე უფრო მკაფიოა.

საშუალო მაჩვენებელი ყველაზე დაუცველებს მალავს. მთლიანობაში გაზომვისას — როგორც ეს ჩვეულებრივ კეთდება — ბევრი მოდელი საკმაოდ კონფიდენციალურად გამოიყურება: პაციენტების უმრავლესობისთვის შეტევა შემთხვევით გამოცნობაზე უკეთ არ მუშაობს. მაგრამ პაციენტების მიხედვით სურათი სხვაგვარია. როგორც კნოლემ კვლევის განცხადებაში აღნიშნა, წინა შეფასებები „მხოლოდ ყველა პაციენტზე საშუალო რისკს ზომავდა. ჩვენ პირველად შევისწავლეთ რისკი ცალკეული პაციენტის

დონეზე — და სურათი სრულიად განსხვავებულია.“ ზოგი ადამიანისთვის შეტევა **თითქმის სრულყოფილად** მუშაობს — ავტორები ამას აფასებენ შეტევის AUC-ით **0.95 ან მეტი** (0.5 მონეტის აგდებას შეესაბამება, 1.0 კი უშეცდომო კლასიფიკაციას) — მაშინაც კი, როდესაც მთელი მონაცემთა ნაკრების საშუალო მაჩვენებელი შემთხვევითობაზე უკეთესი არ ჩანდა.

საშუალო მაჩვენებელი შეიძლება შემთხვევითი გამოცნობის დონესთან ახლოს იყოს, მაშინ როცა ჩანაწერების მცირე „კუდი“ ბევრად უფრო ადვილად იდენტიფიცირდება. ნაშრომის მთავარი აზრია, რომ კონფიდენციალურობის რისკი უნდა შემოწმდეს პაციენტის დონეზე და არა მხოლოდ მთლიანობაში.

დაუცველობა თანაბრად არ ნაწილდება — და ყველაზე მეტად ნაკლებად წარმოდგენილ ჯგუფებს ეხება. თითქმის სრულყოფილი შეტევის მიმართ ყველაზე მოწყვლადი პაციენტები სისტემატურად იყვნენ იმ ჯგუფებიდან, რომლებიც მონაცემებში **ნაკლებად იყვნენ წარმოდგენილი**: ეთნიკური უმცირესობები, იშვიათი დაავადების ფენოტიპები და უჩვეულო გამოსახულებითი მახასიათებლები. ელექტრონული ჩანაწერების მონაცემთა ნაკრებში შავკანიანი პაციენტები ყველაზე მოწყვლად ჩანაწერებს შორის მოსალოდნელზე **31%-ით უფრო ხშირად** ხვდებოდნენ; მამოგრაფიის ნაკრებში კი ავთვისებიანობაზე საეჭვოდ მონიშნული სკანები ამ უკიდურესი რისკის ზონაში შთამბეჭდავი **+1,179%-ით** იყო ზედმეტად წარმოდგენილი. (ეს ორი რიცხვი მაგალითებია და არა სრული სურათი — ნაშრომი მსგავს გადახრას რამდენიმე ჯგუფში აჩვენებს. ისინი აღნიშნავს *ფარდობით ზედმეტ წარმოდგენას* მაღალი რისკის პატარა „კუდში“: რამდენად ხშირად ხვდებიან ეს პაციენტები ყველაზე დაუცველ ჩანაწერებს შორის, და არა იმას, შავკანიანი ან საეჭვო მამოგრაფიის მქონე პაციენტების რა წილია დაუცველი.) მოდელს ასეთი პაციენტების მსგავსი მაგალითები ნაკლებად აქვს,

ამიტომ მათი ჩანაწერები უფრო გამოირჩევა — და სწორედ გამორჩეულობას ეძებს კუთვნილება დასკვნა attack. კონფიდენციალურობის ჩავარდნა შემთხვევითი არ არის; ის კონცენტრირდება მათზე, ვინც მონაცემებში ისედაც ზღვარზეა.

უფრო დიდი მოდელები პრობლემას აუარესებს. თითქმის სრულყოფილი შეტევის მიმართ დაუცველი პაციენტების რაოდენობა მკვეთრად გაიზარდა მოდელის ტევადობასთან ერთად — ერთ დერმატოლოგიურ მონაცემთა ნაკრებში წილი უმცირეს მოდელში პრაქტიკულად ნულიდან ყველაზე დიდ მოდელში დაახლოებით **ათიდან ერთამდე** ავიდა. ავტორები აფრთხილებენ, რომ სამედიცინო AI-მოდელების ზრდასთან და გაძლიერებასთან ერთად — სწორედ ამ მიმართულებით მიდის მთელი სფერო — ეს კონკრეტული რისკიც მატულობს და არა მცირდება.

რიუკერტის შეჯამება მკაცრია: „ეს მისაღები რისკი არ არის. ჯანმრთელობის მონაცემები უაღრესად სენსიტიურია.“

რატომ არის „საშუალოდ უსაფრთხო“ არასწორი ტესტი

დაბალი საშუალო რისკის დანახვისას ადვილია ვიფიქროთ, რომ ყველაფერი წესრიგშია. ნაშრომის ძირითადი გაკვეთილი ისაა, რომ აქ საშუალოს აღება უბრალოდ უხეში შეფასება კი არა, **არასწორი რამის გაზომვაა**.

კონფიდენციალურობის გარანტიას მნიშვნელობა მხოლოდ მაშინ აქვს, თუ ის ყველაზე მაღალი რისკის მქონე ადამიანისთვისაც მოქმედებს და არა მხოლოდ საშუალო პაციენტისთვის. მოდელს, რომელშიც 1,000 პაციენტიდან 999-ის წევრობა პრაქტიკულად ვერ დგინდება, მაგრამ ერთი პაციენტი თითქმის სრულყოფილად იდენტიფიცირდება, ვერ ვუწოდებთ „99.9%-ით კონფიდენციალურს“ იმ აზრით, რომელსაც ამ ერთ ადამიანთან მიმართებით მნიშვნელობა ექნება. და თუ ეს ერთი ადამიანი წინასწარ განჭვრეტად ხშირად იშვიათი დაავადების მქონე ან ეთნიკური უმცირესობის წარმომადგენელია, მაჩვენებელი უბრალოდ არასრული კი არა, ჩუმად დისკრიმინაციულიც ხდება. ის უმრავლესობის უსაფრთხოებას ზომავს და ამას ყველას უსაფრთხოებად აცხადებს.

სწორედ ამ ცვლილებას გვაიძულებს ნაშრომი: კითხვიდან *საშუალოდ რამდენად კონფიდენციალურია ეს მოდელი?* კითხვაზე *ვინ არის ყველაზე დაუცველი პაციენტი და რომელ ჯგუფს ეკუთვნის ის?* ეს სხვადასხვა კითხვებია, და კონფიდენციალურობისთვის სწორედ მეორეა არსებითი.

რას არ ამტკიცებს ეს კვლევა

- ის **არ** ამბობს, რომ სამედიცინო AI უნდა მიტოვდეს. ავტორების მიდგომა რისკის შემცირებაა და არა უკან დახევა: რისკი უნდა გავზომოთ და შევამციროთ, არა

ტექნოლოგიის განვითარება შევწყვიტოთ.

- ის **არ** ნიშნავს, რომ ყველა სამედიცინო მოდელი მონაცემებს ჟონავს ან რომ რომელიმე კონკრეტულ დანერგილ მოდელზე შეტევა უკვე განხორციელდა. ნაშრომი აჩვენებს, რომ *რისკი არსებობს და თანაბრად არ ნაწილდება*, ხოლო სტანდარტული აგრეგირებული მაჩვენებლები მას ვერ ხედავს.
- ის **არ** ნიშნავს, რომ თქვენი ჩანაწერები უკვე გამჟღავნებულია. შეტევას კონკრეტული პირობები სჭირდება — მოდელზე წვდომა, კანდიდატი ჩანაწერი და თავდამსხმელის საკუთარი ინფრასტრუქტურა — და ეს შემთხვევითი შესაძლებლობა არ არის.
- ის **არ** აჩვენებს პაციენტის ჩანაწერის *შინაარსის* გაჟონვას. იჟონება წევრობა — სასწავლო ნაკრებში ჩართვის ფაქტი — რომელიც სხვა მიზეზითაა სახიფათო და არა იმიტომ, რომ მოდელი ჩანაწერს პირდაპირ აბრუნებს.
- ის **არ** ამცირებს განსხვავებებს ერთ სათაურისთვის მოსახერხებელ რიცხვამდე. ძლიერი და განმეორებადი შედეგია *ნიმუში*: აგრეგირებული მაჩვენებლები ინდივიდუალურ რისკს აკნინებს, ხოლო ყველაზე დიდი რისკი ნაკლებად წარმოდგენილ პაციენტებზე მოდის — ეს სურათი მრავალი მონაცემთა ნაკრებისა და ასობით მოდელის მასშტაბით ჩანს.

რამდენად ძლიერია მტკიცებულება?

ეს ემპირიული და მეთოდოლოგიური შედეგია, და საკმაოდ მყარი. ნიმუში — აგრეგირებული კონფიდენციალურობის მაჩვენებლები სისტემატურად ამცირებს ცალკეული პაციენტების რისკს, დარჩენილი რისკი კონცენტრირდება ნაკლებად წარმოდგენილ ჯგუფებზე და მოდელის ტევადობის ზრდასთან ერთად მძიმდება — განმეორდა **შვიდ განსხვავებული ტიპის მონაცემთა ნაკრებში** და თითოეულ მათგანში მრავალ მოდელზე. სწორედ ეს სიგანე ხდის გაზომვით დასკვნას დამაჯერებელს და არა ერთი მონაცემთა ნაკრების თავისებურებას.

ორი მნიშვნელოვანი შენიშვნა. პირველი: ეს არის *რისკის* დემონსტრირება, რომელიც ავტორების მიერ აგებული შეტევებით გაზომეს; ის გვიჩვენებს, რამდენად წარმატებული *შეიძლება* იყოს კომპეტენტური თავდამსხმელი განსაზღვრულ პირობებში და არა იმას, რამდენად ხშირად ხდება ასეთი შეტევა რეალურ სამყაროში. მეორე: თვალში საცემი რიცხვები — ზოგი პაციენტისთვის თითქმის სრულყოფილი წარმატება — განზრახ აღწერს ყველაზე დაუცველ ადამიანებს; სწორედ ესაა კვლევის აზრი. ისინი უნდა წავიკითხოთ როგორც „რისკის კუდი საშუალოზე ბევრად მძიმეა“ და არა როგორც „პაციენტების უმეტესობა დაუცველია“.

შესაბამისი პოზიცია არც პანიკაა და არც ხელის ჩაქნევა: ეს არის ფრთხილი, მრავალ მონაცემთა ნაკრებზე ჩატარებული კვლევა, რომელიც აჩვენებს, რომ სამედიცინო AI-ის კონფიდენციალურობის სერტიფიცირების სტანდარტული გზა საკუთარ ყველაზე ცუდ შემთხვევებს ვერ ხედავს — და ეს შემთხვევები ხშირად იმ პაციენტებზე მოდის, ვისაც შეცდომისთვის ყველაზე ნაკლები სივრცე აქვს.

რატომ არის ეს მნიშვნელოვანი

ორი მიზეზი ამ შედეგს უბრალო ტექნიკურ შენიშვნაზე მნიშვნელოვანს ხდის.

პირველი: ის ცვლის იმას, რას უნდა ნიშნავდეს „კონფიდენციალურობის დამცავი“. თუ მოდელი ქვეყნდება საშუალო კონფიდენციალურობის ქულით, ეს ქულა შეიძლება მართლაც დაბალი იყოს და მაინც მალავდეს პაციენტების მცირე ჯგუფს, რომელთა სასწავლო ნაკრებში ყოფნაც კუთვნილება დასკვნა attack-ით თითქმის სრულყოფილად დგინდება. ნაშრომის პრაქტიკული მოთხოვნა კონკრეტულია: მოდელის გამოშვებამდე კონფიდენციალურობის რისკი შეაფასეთ **თითოეული პაციენტის დონეზე**, აკონტროლეთ ვის აქვს წვდომა დანერგილ მოდელებზე და გამოიყენეთ ისეთი მეთოდები, როგორიცაა **differential privacy** — სწავლებისას მცირე, საგულდაგულოდ დაკალიბრებული ხმაურის დამატება, რომელიც კუთვნილება დასკვნა attack-ს ასუსტებს, თუმცა მოდელის სარგებლიანობის რეალურ და მართვად ფასად. კონფიდენციალურობასა და სარგებლიანობას შორის კომპრომისი რეალურია და უფასო არაა. „საშუალო შევამოწმეთ“ აღარ უნდა ნიშნავდეს, რომ კონფიდენციალურობა შემოწმებულია.

მეორე: ეს კონფიდენციალურობას სამართლიანობასთან აკავშირებს. იგივე ჯგუფები, რომლებიც სამედიცინო მონაცემებში ნაკლებად არიან წარმოდგენილი — და ამიტომ სამედიცინო AI ხშირად ისედაც უარესად ემსახურება — აღმოჩნდნენ იმ ადამიანებადაც, რომელთა კონფიდენციალურობას ეს AI ყველაზე ნაკლებად იცავს. სფერო, რომელიც პირველ უთანასწორობაზე მუშაობს, მეორეს სხვის პრობლემად ვერ ჩათვლის. საუბარი იმავე ადამიანებზეა.

სამედიცინო AI-ის კონფიდენციალურობის დამამშვიდებელი ამბავი — *გავწომეთ, საშუალო დაბალია, ყველაფერი კარგადაა* — სწორედ ის ამბავია, რომელსაც ეს ნაშრომი შლის. არა იმისთვის, რომ ტექნოლოგია ვინმეს შეაშინოს, არამედ იმისთვის, რომ სტანდარტი იქ გადავიდეს, სადაც უნდა იყოს: დაპირება მხოლოდ მაშინ შეგიძლიათ დადებულად ჩათვალოთ, თუ შეამოწმეთ, სრულდება თუ არა ის ყველაზე მეტად დასაზიანებელი ადამიანის მიმართაც.

მოკლე შეჯამება

კუთვნილება დასკვნა attack ცდილობს დაადგინოს, იყო თუ არა კონკრეტული ადამიანის ჩანაწერი AI-მოდელის სასწავლო მონაცემებში — და რადგან წევრობამ შეიძლება თავისთავად გაამჟღავნოს სენსიტიური ინფორმაცია, მაგალითად კონკრეტული დაავადების არსებობა, ეს კონფიდენციალურობის ნამდვილი გაჟონვაა მაშინაც კი, როცა თავად ჩანაწერი არასოდეს გამოჩნდება. წინა კვლევები ზომავდა, რამდენად ხშირად მუშაობს ასეთი შეტევა მონაცემთა ნაკრებზე *საშუალოდ*. ამ კვლევამ კი რისკი გაზომა *თითოეული პაციენტისთვის*, შვიდ სამედიცინო მონაცემთა ნაკრებში — გამოსახულებები, ECG და ელექტრონული ჩანაწერები — და თითოეულზე მრავალ მოდელში. შედეგი: აგრეგირებული მაჩვენებლები რისკს მნიშვნელოვნად აკნინებს; ზოგი ადამიანის წევრობა თითქმის სრულყოფილად დგინდება (attack AUC $\geq$ 0.95)

მაშინაც კი, როცა მთელი ნაკრების საშუალო უსაფრთხოდ გამოიყურება; ყველაზე დაუცველები სისტემატურად ნაკლებად წარმოდგენილი ჯგუფების წარმომადგენლები არიან — ეთნიკური უმცირესობები, იშვიათი დაავადებები, უჩვეულო გამოსახულებები — და პრობლემა მოდელის ზომასთან ერთად იზრდება. ავტორები სამედიცინო AI-ის მიტოვებას არ ითხოვენ; ისინი გვთავაზობენ კონფიდენციალურობის პაციენტის დონეზე გაზომვას, მოდელზე წვდომის კონტროლს და differential privacy-ის გამოყენებას. მთავარი დასკვნა: „საშუალოდ კონფიდენციალური“ კონფიდენციალურობის გარანტია არ არის, და ეს მიდგომა ყველაზე ხშირად სწორედ ისედაც ნაკლებად დაცულ პაციენტებს ღალატობს.

პირდაპირი შემოწმება

რას აჩვენებს ნაშრომი: შვიდ სამედიცინო მონაცემთა ნაკრებში და თითოეულ მათგანზე მრავალ მოდელში, პაციენტის დონეზე membership-დასკვნა რისკი ზოგი ადამიანისთვის ბევრად უფრო მაღალია, ვიდრე აგრეგირებული მაჩვენებლები მიანიშნებს — შეტევა ზოგ შემთხვევაში თითქმის სრულყოფილია (AUC $\geq$ 0.95). დარჩენილი რისკი არაპროპორციულად ნაკლებად წარმოდგენილ ჯგუფებზე მოდის და მოდელის ტევადობის ზრდასთან ერთად მატულობს.

რა არის შესაძლებელი, მაგრამ დაუმტკიცებელი: რომ რეალური თავდამსხმელები ამას უკვე იყენებენ დანერგილი კლინიკური მოდელების წინააღმდეგ; კვლევა აღწერილ პირობებში აჩვენებს *შესაძლებლობას და მის განაწილებას*, არა რეალურ სამყაროში შეტევების სიხშირეს.

რას არ აჩვენებს: რომ სამედიცინო AI უნდა მიტოვდეს; რომ ყველა მოდელი ჟონავს ან რომელიმე კონკრეტული დანერგილი მოდელი გატეხილია; რომ გამჟღავნებულია პაციენტის ჩანაწერის *შინაარსი* და არა მხოლოდ წევრობა; ან რომ განსხვავება ერთი უნივერსალური რიცხვით შეიძლება გამოიხატოს — მყარი შედეგი არის ნიმუში და არა ერთი მაჩვენებელი.

მთავარი შეზღუდვები: კვლევა ზომავს ყველაზე ცუდი შემთხვევის რისკს ავტორების მიერ განხორციელებული შეტევებით და არა რეალურად დაფიქსირებულ ინციდენტებს; ყველაზე შთამბეჭდავი რიცხვები განზრახ აღწერს ყველაზე დაუცველ ადამიანებს; ჯგუფებს შორის ზუსტი სხვაობები წარმოდგენილია სხვადასხვა მონაცემთა ნაკრებში განმეორებადი ნიმუშის სახით და არა ერთ რიცხვად.

რამდენად უნდა ენდოს ამას ზოგადი მკითხველი? მაღალი ნდობაა გამართლებული იმაში, რომ აგრეგირებული კონფიდენციალურობის მაჩვენებლები ინდივიდუალურ რისკს აკნინებს, დარჩენილი რისკი ნაკლებად წარმოდგენილ პაციენტებზე კონცენტრირდება და მოდელის ზომასთან ერთად იზრდება — ეს მრავალი მონაცემთა ნაკრებისა და მოდელის მასშტაბითაა ნაჩვენები. ზომიერი ნდობაა საჭირო რეალურ სამყაროში ასეთი შეტევების სიხშირის შეფასებისას, რადგან ეს კვლევა მას არ ზომავს. უსაფრთხო წაკითხვა ასეთია: არა „სამედიცინო AI თქვენს მონაცემებს ჟონავს“ და არც „კონფიდენციალურო უკვე მოგვარებულია“, არამედ „სტანდარტული კონფიდენციალურობის შემოწმება

საკუთარ ყველაზე ცუდ შემთხვევებს ვერ ხედავს, და ეს შემთხვევები ყველაზე დაუცველ პაციენტებზე მოდის — ამიტომ შემოწმების მეთოდი უნდა შეიცვალოს.“

წყაროები

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