

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

AI medis bisa privat secara rata-rata tetapi tetap mengekspos pasien tertentu — terutama yang kurang terwakili

Model AI medis yang disebut "menjaga privasi" biasanya bertumpu pada satu angka rata-rata. Studi ini berargumen bahwa itu tes yang salah. Dengan mempelajari membership inference attack — yang mengungkap apakah rekam seseorang berada dalam data pelatihan model dan dengan demikian dapat membocorkan bahwa ia memiliki penyakit tertentu — para penulis mengukur risiko per pasien, bukan secara agregat, pada tujuh dataset medis (pencitraan, ECG, rekam elektronik) dan banyak model. Polanya: model yang tampak aman secara rata-rata masih dapat membuat penyerang mengidentifikasi individu tertentu hampir sempurna (attack AUC ≥ 0,95); yang terekspos secara sistematis berasal dari kelompok kurang terwakili (etnis minoritas, penyakit langka, pencitraan tidak biasa); dan masalah memburuk saat model membesar. Para penulis tidak mengatakan tinggalkan AI medis — mereka meminta privasi diukur per pasien, akses model dikendalikan, dan differential privacy digunakan. Inti yang tidak nyaman: "privat secara rata-rata" bukan jaminan privasi, dan gagal terutama pada pasien yang sudah paling kurang terlindungi.

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

Jaminan privasi adalah janji kepada seseorang, bukan kepada rata-rata

Ketika model AI medis disebut "menjaga privasi", klaim itu biasanya bertumpu pada satu angka: di antara semua pasien yang datanya dipakai untuk melatih model, peluang rata-rata bahwa keanggotaan seseorang dalam data latih dapat diinferensikan rendah. Kedengarannya menenangkan. Poin tenang tetapi tidak nyaman dari makalah ini adalah: itu angka yang salah.

Privasi bukanlah rata-rata. Ia adalah janji kepada setiap individu — bahwa keberadaannya dalam set pelatihan tidak akan berbalik mengekspos dirinya. Sebuah rata-rata dapat menepati janji itu bagi hampir semua orang sambil melanggarnya sepenuhnya bagi segelintir orang. Para peneliti mengukur risiko privasi satu pasien demi satu pasien, pada resolusi yang belum pernah dipakai sebelumnya, dan menemukan tepat hal itu: model yang tampak aman secara agregat dapat membocorkan keanggotaan individu tertentu hampir sempurna — dan individu yang bocor secara tidak proporsional adalah mereka yang sejak awal paling kurang terlindungi.

Apa yang dilakukan para penulis

Tim — dipimpin Moritz Knolle, bersama Daniel Rückert (keduanya Technical University of Munich) dan Georg Kaissis (Hasso Plattner Institute), serta kolega dari Imperial College London — mengambil ancaman privasi yang dikenal sebagai **serangan inferensi keanggotaan** dan mengubah pertanyaannya. Alih-alih bertanya "secara rata-rata, seberapa sering serangan ini berhasil pada seluruh kumpulan data?", mereka bertanya "untuk *pasien ini secara khusus*, seberapa terekspos ia?"

Mereka menjalankan analisis per pasien itu pada **tujuh kumpulan data medis mapan**, yang mencakup jenis data sangat berbeda — pencitraan medis, elektrokardiogram, dan rekam kesehatan elektronik — serta 200 model untuk masing-masing kumpulan data. Untuk setiap pasien pada setiap model, mereka memperkirakan seberapa yakin penyerang dapat menentukan apakah rekam orang tersebut menjadi bagian data pelatihan, lalu memecah hasil berdasarkan kelompok: status penyakit, ras yang dilaporkan sendiri, asuransi, jenis kelamin, dan protokol pencitraan.

> **Apa sebenarnya serangan inferensi keanggotaan?**
>
> Kebocoran di sini lebih halus daripada "model mengeluarkan rekam medis Anda". Yang bocor adalah **status keanggotaan** — fakta bahwa data Anda berada dalam kumpulan pelatihan. Mulailah dari apa yang biasanya dilakukan model seperti ini. Anda memasukkan citra — misalnya foto rontgen dada — lalu model memberikan probabilitas: peluang 78% pneumonia, beserta pembacaan untuk kardiomegali, edema, konsolidasi. Jawaban diagnostik sehari-hari itu adalah *satu-satunya* hal yang digunakan serangan. Tidak ada API rahasia atau akses aneh. Kelemahan yang dimanfaatkan adalah ini: model biasanya sedikit **lebih yakin** pada contoh persis yang dipakai untuk melatihnya dibanding contoh yang belum pernah dilihat. Bayangkan siswa yang diam-diam pernah melihat soal ujian — pada soal yang sudah dilatih,

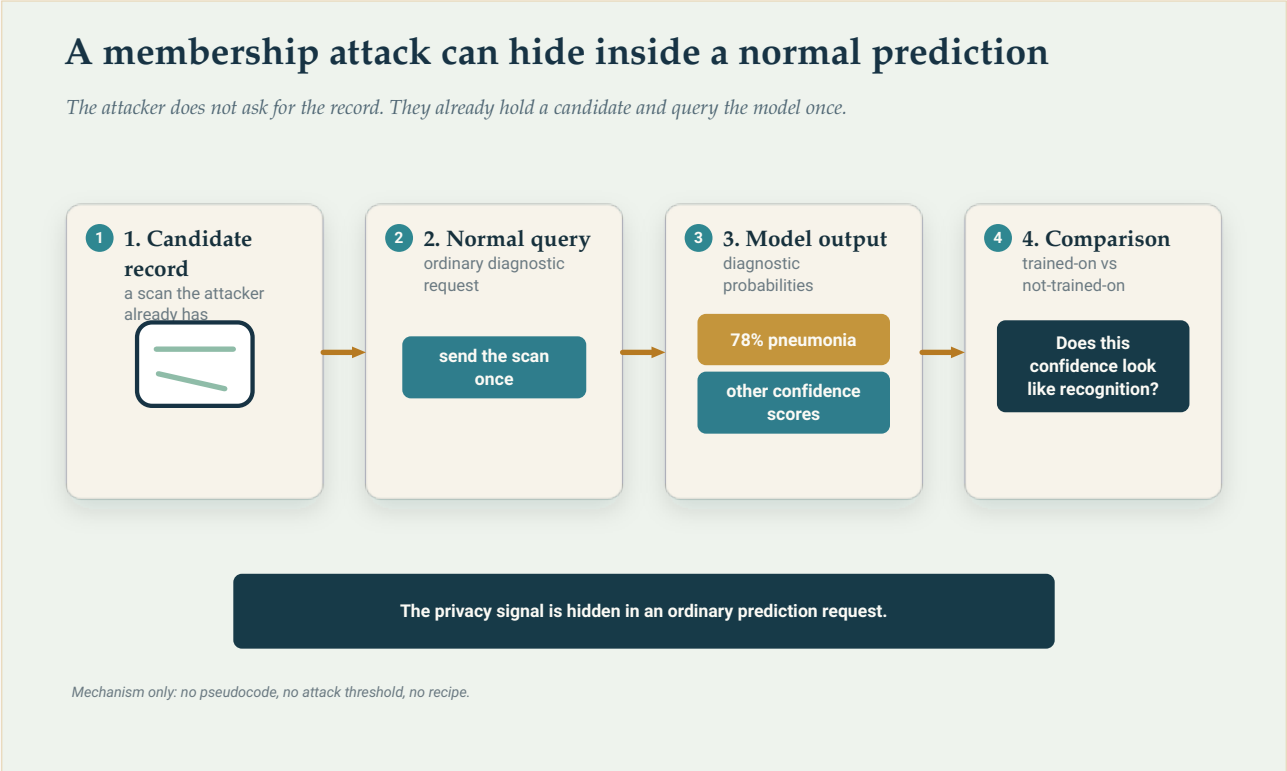
jawabannya keluar sedikit terlalu cepat dan terlalu yakin. Menyimpulkan dari petunjuk itu apakah suatu rekam pernah masuk ke data pelatihan adalah yang disebut **inferensi keanggotaan** (*membership inference*).

Serangannya, langkah demi langkah: seseorang ingin tahu apakah citra orang tertentu dipakai melatih model. Mereka **tidak** meminta model mengeluarkan rekam tersebut — mereka sudah memiliki citra kandidat (milik orang itu sendiri, atau salinan yang sangat mirip). Citra dimasukkan sekali sebagai permintaan diagnosis biasa, lalu tingkat keyakinan model dicatat. Kemudian keyakinan itu dibandingkan dengan pola model yang *pernah* dilatih pada citra dan model yang *tidak pernah* melihatnya — perbandingan yang dapat dipelajari secara relatif murah dengan melatih model referensi sendiri pada komputer biasa tanpa perangkat keras khusus. Jika model sebenarnya luar biasa yakin pada citra tersebut — seolah mengenalinya — itu menjadi bukti kuat bahwa scan ada dalam data pelatihan.

Bagian yang mengganggu adalah permintaan tersebut sama sekali tidak terlihat seperti serangan: ia sama dengan kueri diagnostik yang akan dilakukan klinisi, dan sinyal privasinya tersembunyi di dalam prediksi biasa. Itulah mengapa risikonya konkret — meskipun sejauh ini ditunjukkan di bawah asumsi laboratorium yang dinyatakan, bukan ditemukan dalam serangan dunia nyata.

Mengapa status keanggotaan penting jika isi rekam tidak pernah bocor? Karena *keanggotaan itu sendiri adalah fakta*. Jika suatu model dilatih pada pasien yang menerima imunoterapi kanker tertentu, memastikan bahwa rekam Anda ada di dalamnya dapat mengungkap bahwa Anda kemungkinan menderita kanker tersebut — jenis informasi yang tidak seharusnya dapat diinferensikan perusahaan asuransi atau pemberi kerja. Isinya tetap tertutup; fakta bahwa Anda termasuklah yang lolos.

Para penulis menyatakan asumsi ini dengan jelas — akses ke prediksi biasa model, rekam kandidat, dan model referensi milik penyerang — bukan sebagai panduan serangan, melainkan agar ancamannya dapat dipikirkan secara konkret alih-alih diabaikan.



Serangan tidak membutuhkan API privasi khusus. Kueri diagnostik biasa dapat membocorkan sinyal keanggotaan jika model luar biasa yakin pada rekam yang pernah dilihat sebelumnya.

Original Aurora diagram — The Clean Paper CC BY 4.0

Apa yang mereka temukan

Tiga temuan, masing-masing lebih tajam daripada sebelumnya.

Rata-rata menyembunyikan orang yang terekspos. Jika diukur secara agregat — cara yang biasa digunakan — banyak model tampak cukup aman: serangan tidak lebih baik daripada tebakan acak bagi kebanyakan pasien. Pandangan per pasien bercerita lain. Seperti dikatakan Knolle dalam pengu-muman studi, penilaian sebelumnya "hanya pernah mengukur risiko rata-rata di seluruh pasien. Kami memeriksa risiko pada tingkat pasien individual untuk pertama kalinya — dan gambarnya sangat berbeda." Untuk beberapa individu, serangan berhasil **hampir sempurna** — diukur dengan **AUC serangan 0,95 atau lebih tinggi** (0,5 setara lempar koin, 1,0 sempurna) — bahkan ketika rata-rata seluruh kumpulan data tampak tidak lebih baik daripada tebakan acak.



Rata-rata dapat berada dekat tingkat kebetulan sementara sebagian kecil rekam pada ekor distribusi tetap jauh lebih terekspos. Poin makalah adalah privasi harus diperiksa pada tingkat pasien, bukan hanya secara agregat.

Original Aurora diagram — The Clean Paper CC BY 4.0

Eksposur tidak merata — dan mengenai kelompok yang kurang terwakili. Pasien yang paling rentan terhadap serangan hampir sempurna secara sistematis berasal dari kelompok yang **kurang terwakili dalam data**: etnis minoritas, fenotipe penyakit langka, karakteristik pencitraan tidak umum. Dalam kumpulan data rekam elektronik, pasien kulit hitam muncul **31% lebih sering** daripada yang diharapkan di antara rekam paling rentan; pada kumpulan data mamografi, scan yang ditandai mencurigakan terhadap keganasan terwakili berlebihan di kelompok paling berisiko itu sebesar **+1.179%.** (Dua angka ini adalah contoh, bukan seluruh gambaran — makalah melaporkan pola serupa pada beberapa kelompok — dan keduanya adalah *keterwakilan berlebih relatif* dalam ekor kecil berisiko ekstrem: seberapa lebih sering pasien ini muncul di antara rekam paling terekspos, bukan persentase seluruh pasien kulit hitam atau mamografi mencurigakan yang terekspos.) Model memiliki lebih sedikit contoh serupa untuk "membaurkan" pasien ini, sehingga rekam mereka menonjol — dan menonjol adalah tepat hal yang dideteksi serangan inferensi keanggotaan. Kegagalan privasi bukan acak; ia terkonsentrasi pada orang yang memang sudah berada di pinggiran.

Model lebih besar memperburuknya. Jumlah pasien yang terekspos terhadap serangan hampir sempurna meningkat tajam dengan kapasitas model — dalam satu kumpulan data dermatologi, proporsinya naik dari hampir nol pada model terkecil menjadi sekitar **satu dari sepuluh** pada model terbesar. Seiring model AI medis tumbuh semakin besar dan mampu — arah yang sedang ditempuh seluruh bidang — risiko spesifik ini, menurut para penulis, menjadi semakin berat, bukan berkurang.

Ringkasan Rückert lugas: "Ini bukan risiko yang dapat ditoleransi. Data kesehatan sangat sensitif."

Mengapa “aman secara rata-rata” adalah tes yang salah

Godaan mudahnya adalah membaca risiko rata-rata rendah sebagai sertifikat aman. Pelajaran utama makalah ini adalah bahwa rata-rata di sini bukan sekadar kurang presisi — ia mengukur hal yang salah.

Jaminan privasi hanya bermakna jika berlaku pada orang dengan risiko tertinggi, bukan orang di tengah distribusi. Model yang membuat 999 dari 1.000 pasien tidak dapat dikenali tetapi satu pasien dapat diidentifikasi hampir sempurna bukan “99,9% privat” dalam arti yang relevan bagi orang tersebut — dan jika orang itu secara dapat diprediksi adalah pasien penyakit langka atau pasien dari etnis minoritas, metriknya bukan hanya tidak lengkap, melainkan diam-diam diskriminatif. Ia melaporkan keamanan mayoritas lalu menyebutnya keamanan untuk semua.

Itulah pergeseran yang dipaksakan makalah: dari *seberapa privat model ini secara rata-rata?* menjadi *siapa pasien yang paling terekspos, dan siapa mereka?* Itu dua pertanyaan berbeda, dan hanya yang kedua benar-benar pertanyaan privasi.

Apa yang *tidak* dibuktikan — atau diklaim

- Studi ini **tidak** mengatakan AI medis harus ditinggalkan. Kerangka penulis adalah mitigasi, bukan mundur: ukur dan perbaiki risiko, jangan berhenti membangun.
- Studi ini **tidak** berarti setiap model medis bocor, atau model tertentu yang sudah digunakan pernah diserang. Ia menunjukkan *risiko ada dan distribusinya tidak merata,* serta metrik agregat standar tidak melihatnya.
- Studi ini **tidak** berarti rekam Anda sudah terekspos. Serangan membutuhkan kondisi tertentu — akses ke model, rekam kandidat, dan infrastruktur referensi milik penyerang — bukan kemampuan kasual.
- Studi ini **tidak** menunjukkan *isi* rekam pasien bocor. Yang bocor adalah status keanggotaan — fakta keterlibatan — yang berbahaya karena alasan berbeda, bukan karena rekam lengkap dikeluarkan.
- Studi ini **tidak** mereduksi ketimpangan ke satu angka judul utama. Hasil kuat dan dapat direproduksi adalah *polanya* — metrik agregat mengecilkan risiko individu, dan pasien yang kurang terwakili menanggung bagian terbesar — pada banyak kumpulan data dan ratusan model.

Seberapa kuat buktinya?

Ini hasil empiris dan metodologis, dan cukup kokoh. Polanya — metrik privasi agregat secara sistematis mengecilkan risiko per pasien, sisa risiko terkonsentrasi pada kelompok kurang terwakili, dan memburuk saat kapasitas model naik — bertahan pada **tujuh kumpulan data dengan tipe data berbeda** dan banyak model untuk tiap kumpulan data. Cakupan itu tepat yang membuat klaim pengukuran lebih kredibel daripada artefak satu kumpulan data.

Ada dua catatan jujur. Pertama, ini demonstrasi *risiko*, diukur dengan menjalankan serangan yang dibangun para penulis sendiri; ia menggambarkan seberapa baik penyerang yang mampu *dapat* bekerja di bawah asumsi tertentu, bukan seberapa sering serangan nyata terjadi. Kedua, angka dramatis — keberhasilan hampir sempurna pada sebagian pasien — memang menggambarkan individu dengan risiko terburuk *secara sengaja*; itulah intinya, dan harus dibaca sebagai "ekornya jauh lebih berat daripada yang disiratkan rata-rata", bukan "kebanyakan pasien terekspos".

Sikap yang tepat bukan alarm maupun penolakan: ini studi berbasis banyak kumpulan data yang hati-hati, menunjukkan bahwa cara standar kita mensertifikasi privasi AI medis buta terhadap kasus terburuknya sendiri — dan kasus terburuk itu mengenai pasien yang paling sedikit memiliki margin perlindungan.

Mengapa ini penting

Dua hal membuat temuan ini lebih dari catatan teknis.

Pertama, ia mengubah apa yang boleh dimaksud dengan "menjaga privasi". Jika model dirilis dengan satu skor privasi rata-rata, skor itu dapat benar-benar rendah dan tetap menyembunyikan sekelompok pasien yang hampir dapat dikenali dengan sempurna oleh serangan inferensi keanggotaan. Tuntutan praktis makalah konkret: nilai risiko privasi **per pasien** sebelum rilis, kendalikan siapa yang dapat mengakses model yang diterapkan, dan gunakan teknik seperti **privasi diferensial** — derau kecil yang dikalibrasi secara hati-hati selama pelatihan untuk menumpulkan serangan inferensi keanggotaan, dengan biaya nyata dan terkelola terhadap kegunaan model (kompromi antara privasi dan kegunaan itu nyata, bukan tanpa biaya). "Kami sudah memeriksa rata-ratanya" seharusnya berhenti dianggap sebagai sudah memeriksa privasi.

Kedua, makalah ini menghubungkan privasi dengan **keadilan** (*fairness*). Kelompok yang sama yang kurang terwakili dalam data medis — dan karena itu sudah lebih buruk dilayani AI medis — ternyata juga kelompok yang privasinya paling sedikit dilindungi oleh AI tersebut. Bidang yang bekerja keras pada ketimpangan pertama tidak dapat memperlakukan yang kedua sebagai urusan departemen lain. Orangnya sama.

Cerita yang menenangkan tentang privasi AI medis — *kami sudah mengukurnya, rata-ratanya rendah, jadi aman* — adalah cerita yang dibongkar makalah ini. Bukan untuk menakut-nakuti orang dari teknologi, tetapi untuk memindahkan standar ke tempat yang seharusnya: sebuah janji yang hanya boleh diklaim dipenuhi jika sudah diperiksa pada orang yang paling mungkin dirugikan.

Singkatnya

Serangan inferensi keanggotaan mencoba menentukan apakah rekam seseorang pernah berada dalam data pelatihan model AI — dan karena status keanggotaan itu sendiri dapat mengungkap informasi (misalnya bahwa seseorang memiliki penyakit tertentu), itu kebocoran privasi nyata

bahkan ketika isi rekam tidak pernah muncul. Penelitian sebelumnya mengukur seberapa sering serangan berhasil *secara rata-rata* di seluruh kumpulan data. Studi ini mengukurnya *per pasien*, pada tujuh kumpulan data medis (pencitraan, ECG, rekam elektronik) dan banyak model, lalu menemukan bahwa metrik agregat sangat mengecilkan risiko: sebagian individu dapat diidentifikasi hampir sempurna (AUC serangan $\geq$ 0,95) bahkan ketika rata-rata kumpulan data tampak aman; individu yang paling rentan secara sistematis berasal dari kelompok yang kurang terwakili (etnis minoritas, penyakit langka, pencitraan tidak biasa); dan masalah membesar bersama ukuran model. Para penulis tidak meminta AI medis ditinggalkan — mereka meminta privasi diukur pada tingkat individu, akses model dikendalikan, dan privasi diferensial digunakan. Intinya: "privat secara rata-rata" bukan jaminan privasi, dan orang yang paling gagal dilindunginya biasanya justru yang sudah paling kurang terlindungi.

Penilaian tanpa berlebihan

Apa yang ditunjukkan makalah: Pada tujuh kumpulan data medis dan banyak model, risiko inferensi keanggotaan per pasien jauh lebih tinggi bagi sebagian individu daripada yang disiratkan metrik agregat — hingga keberhasilan serangan hampir sempurna (AUC $\geq$ 0,95) — dan sisa risiko itu secara tidak proporsional mengenai kelompok kurang terwakili serta meningkat bersama kapasitas model.

Apa yang masuk akal tetapi belum terbukti: Bahwa penyerang dunia nyata sudah mengeksploitasi hal ini terhadap model klinis yang digunakan; studi menunjukkan *kemampuan dan distribusinya* di bawah asumsi tertentu, bukan frekuensi serangan di alam liar.

Apa yang tidak ditunjukkan: Bahwa AI medis harus ditinggalkan; bahwa setiap model bocor atau model tertentu telah ditembus; bahwa *isi* rekam pasien, bukan status keanggotaan, terekspos; satu angka ketimpangan tunggal — hasil yang kokoh adalah pola, bukan satu angka.

Keterbatasan utama: Mengukur risiko kasus terburuk menggunakan serangan yang dibangun penulis, bukan insiden yang diamati; angka dramatis memang menggambarkan individu paling terekspos secara sengaja; dan ketimpangan spesifik per kelompok ditunjukkan sebagai pola konsisten lintas kumpulan data alih-alih dipadatkan menjadi satu nilai.

Seberapa besar keyakinan yang sebaiknya dimiliki pembaca umum? Tinggi bahwa metrik privasi agregat mengecilkan risiko individu, bahwa sisa risiko terkonsentrasi pada pasien kurang terwakili, dan bahwa masalah memburuk dengan ukuran model — semuanya ditunjukkan di banyak kumpulan data dan model. Sedang mengenai frekuensi serangan dunia nyata, yang tidak diukur studi ini. Pembacaan yang aman bukan "AI medis membocorkan data Anda" dan bukan "privasi sudah selesai", melainkan "pemeriksaan privasi standar buta terhadap kasus terburuknya sendiri, dan kasus itu mengenai pasien paling rentan — jadi pemeriksaannya harus berubah."

Sumber

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