

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

AI perubatan boleh kelihatan peribadi secara purata tetapi tetap mendedahkan pesakit tertentu — paling teruk yang kurang diwakili

Model AI perubatan yang dipanggil 'memelihara privasi' biasanya bergantung pada satu nombor purata. Kajian ini berhujah bahawa nombor itu ialah ujian yang salah. Dengan mengkaji serangan inferens keahlian — yang mendedahkan sama ada rekod seseorang tertentu berada dalam data latihan model, lalu boleh membocorkan bahawa mereka mempunyai penyakit tertentu — para penulis mengukur risiko bagi setiap pesakit, bukan secara agregat, merentas tujuh set data perubatan (pengimejan, ECG, rekod elektronik) dan banyak model bagi setiap set. Coraknya: model yang kelihatan selamat secara purata masih boleh membenarkan penyerang mengenal pasti individu tertentu hampir sempurna (AUC serangan ≥ 0.95); yang terdedah secara sistematik datang daripada kumpulan kurang diwakili (etnik minoriti, penyakit jarang, pengimejan luar biasa); dan masalah bertambah buruk apabila model membesar. Para penulis tidak berkata tinggalkan AI perubatan — mereka berkata ukur privasi bagi setiap pesakit, kawal akses kepada model, dan gunakan privasi pembezaan. Inti yang tidak selesa: 'peribadi secara purata' bukan jaminan privasi, dan ia gagal pada pesakit yang memang paling kurang dilindungi.

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 ialah janji kepada seseorang, bukan kepada purata

Apabila model AI perubatan dipanggil “memelihara privasi”, dakwaan itu biasanya bergantung pada satu nombor: merentas semua pesakit yang datanya digunakan untuk melatih model, purata kemungkinan keahlian mana-mana individu boleh disimpulkan adalah rendah. Kedengarannya meyakinkan. Inti senyap dan tidak selesa kertas ini ialah bahawa nombor itu nombor yang salah.

Privasi bukan purata. Ia ialah janji yang dibuat kepada setiap individu — bahawa berada dalam set latihan tidak akan kembali untuk mendedahkan mereka. Dan purata boleh menunaikan janji itu untuk hampir semua orang sambil melanggarnya sepenuhnya bagi segelintir. Para penyelidik mengukur risiko privasi satu pesakit demi satu, pada resolusi yang belum pernah digunakan sebelum ini, dan menemui tepat perkara itu: model yang kelihatan selamat secara agregat boleh membocorkan keahlian individu tertentu hampir sempurna — dan individu yang dibocorkan secara tidak seimbang ialah mereka yang sudah paling kurang dilindungi.

Apa yang dilakukan oleh para penulis

Pasukan — diketuai Moritz Knolle, bersama Daniel Rückert (kedua-duanya Technical University of Munich) dan Georg Kaissis (Hasso Plattner Institute), serta rakan-rakan di Imperial College London — mengambil ancaman privasi terkenal yang dipanggil **serangan inferens keahlian** dan menukar soalannya. Daripada bertanya “secara purata, berapa kerap serangan ini berjaya di seluruh set data?”, mereka bertanya “bagi *pesakit tertentu ini*, sejauh mana dia terdedah?”

Mereka menjalankan analisis setiap pesakit itu merentas **tujuh set data perubatan mapan**, meliputi jenis data yang sangat berbeza — pengimejan perubatan, elektrokardiogram dan rekod kesihatan elektronik — dan, bagi setiap set, 200 model. Untuk setiap pesakit dalam setiap model, mereka menganggarkan sejauh mana yakin penyerang boleh menentukan sama ada rekod orang itu pernah menjadi sebahagian daripada data latihan, kemudian memecahkan hasil mengikut kumpulan: status penyakit, ras yang dilaporkan sendiri, insurans, jantina dan protokol pengimejan.

> **Apa sebenarnya serangan inferens keahlian**
>
> Kebocoran di sini lebih halus daripada “model mengeluarkan rekod anda.” Ia tentang *keahlian* — fakta semata-mata bahawa data anda berada dalam set latihan.
>
> Mulakan dengan apa yang biasanya dilakukan salah satu model ini. Anda memberinya imbasan — contohnya X-ray dada — dan ia memberikan kebarangkalian: 78% kemungkinan pneumonia, bersama bacaan bagi kardiomegali, edema, konsolidasi. Jawapan diagnostik biasa itu ialah *satu-satunya* perkara yang digunakan serangan. Tiada apa yang eksotik.
>
> Kelemahan yang digunakan ialah ini: model biasanya sedikit **lebih yakin** terhadap contoh tepat yang pernah digunakan untuk melatihnya berbanding contoh yang tidak pernah dilihat. Bayangkan pelajar yang secara rahsia melihat peperiksaan terlebih dahulu — pada soalan yang

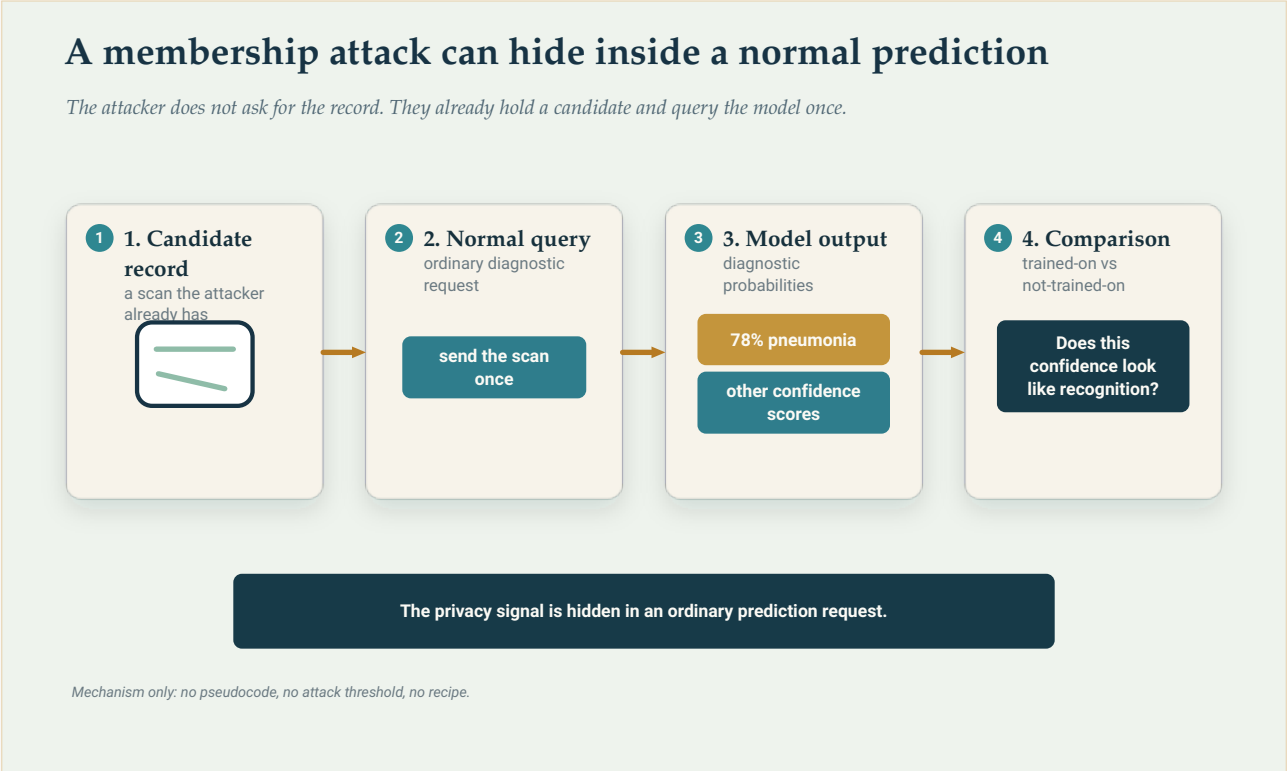
pernah mereka latih, jawapan datang sedikit terlalu cepat, sedikit terlalu yakin. Menentukan, daripada petunjuk itu, sama ada rekod tertentu ialah salah satu contoh yang pernah dipelajari model ialah maksud "inferens keahlian."

Jadi serangannya begini, langkah demi langkah. Seseorang mahu tahu sama ada imbasan individu tertentu digunakan untuk melatih model. Mereka **tidak** meminta model mengeluarkan rekod itu — mereka sudah memiliki imbasan calon (imbasan orang itu sendiri, atau salinan yang hampir sama). Mereka menghantarnya sekali sebagai permintaan diagnostik biasa dan mencatat sejauh mana yakin jawapan. Kemudian mereka memeriksa sama ada keyakinan itu lebih menyerupai model yang *pernah* dilatih pada imbasan tersebut atau model yang *tidak pernah* — perbandingan yang boleh dibuat dengan murah dengan melatih model pengganti sendiri ("model rujukan") untuk mempelajari bagaimana rupa keyakinan berlebihan yang menjadi petunjuk, pada komputer biasa tanpa perkakasan khas. Jika model sebenar luar biasa yakin terhadap imbasan itu — seolah-olah mengenalinya — itu bukti kuat bahawa imbasan tersebut berada dalam data latihan.

Bahagian yang mengganggu ialah tiada apa pada permintaan itu kelihatan seperti serangan: ia pertanyaan diagnostik yang sama yang akan dibuat klinisian, dan isyarat privasi tersembunyi dalam ramalan biasa. Itulah sebab risikonya konkrit — walaupun setakat ini ia ditunjukkan di bawah andaian makmal yang dinyatakan, bukan dikesan secara nyata di dunia sebenar.

Mengapa keahlian penting jika rekod itu sendiri tidak pernah bocor? Kerana *keahlian ialah fakta*. Jika model dilatih pada pesakit yang menerima imunoterapi kanser tertentu, mengesahkan bahawa rekod anda ada di dalamnya mendedahkan bahawa anda mungkin menghidap kanser itu — jenis perkara yang sepatutnya tidak pernah boleh disimpulkan oleh syarikat insurans atau majikan. Kandungan kekal tertutup; fakta bahawa anda tergolong dalam set itulah yang terlepas.

Para penulis menyatakan andaian ini dengan jelas — akses kepada ramalan biasa model, rekod calon, dan model rujukan milik penyerang sendiri — bukan sebagai panduan serangan, tetapi supaya ancaman boleh dinilai secara nyata dan bukannya diketepikan secara kabur.



Serangan tidak memerlukan API privasi khas. Pertanyaan diagnostik biasa boleh membocorkan isyarat keahlian jika model luar biasa yakin terhadap rekod yang pernah dilihatnya.

Original Aurora diagram — The Clean Paper CC BY 4.0

Apa yang mereka temui

Tiga penemuan, setiap satu lebih tajam daripada sebelumnya.

Purata menyembunyikan individu yang terdedah. Diukur secara agregat — cara biasa — banyak model ini kelihatan meyakinkan dari segi privasi: serangan tidak lebih baik daripada tekaan rawak bagi kebanyakan pesakit. Pandangan setiap pesakit memberikan cerita berbeza. Seperti yang dikatakan Knolle dalam pengumuman kajian, penilaian terdahulu "hanya pernah mengukur risiko purata merentas semua pesakit. Kami memeriksa risiko pada peringkat pesakit individu buat kali pertama — dan gambarnya sangat berbeza." Bagi sesetengah individu, serangan berjaya **hampir sempurna** — penulis mengukurnya sebagai AUC serangan **0.95 atau lebih tinggi** (0.5 sama seperti lambungan syiling, 1.0 sempurna) — walaupun purata keseluruhan set data kelihatan tidak lebih baik daripada tekaan rawak.



Purata boleh berada berhampiran tahap tekaan rawak sementara satu ekor kecil rekod kekal jauh lebih terdedah. Inti kertas ialah privasi perlu diperiksa pada peringkat pesakit, bukan hanya secara agregat.

Original Aurora diagram — The Clean Paper CC BY 4.0

Pendedahan itu tidak sama rata — dan menimpa yang kurang diwakili. Pesakit paling terdedah kepada serangan hampir sempurna secara sistematik datang daripada kumpulan yang **kurang diwakili dalam data**: etnik minoriti, fenotip penyakit jarang, ciri pengimejan luar biasa. Dalam set data rekod elektronik, pesakit Kulit Hitam muncul **31% lebih kerap** daripada yang dijangka dalam kalangan rekod paling terdedah; dalam set mamografi, imbasan yang ditandai mencurigakan bagi malignansi terlebih diwakili dalam zon bahaya itu sebanyak **+1,179%**. (Dua angka ini contoh, bukan keseluruhan gambaran — kertas melaporkan kecondongan sama merentas beberapa kumpulan — dan ia ialah *lebihan perwakilan relatif* dalam ekor kecil berisiko ekstrem: berapa lebih kerap pesakit ini muncul antara rekod paling terdedah, bukan pecahan semua pesakit Kulit Hitam atau mamografi mencurigakan yang terdedah.) Model mempunyai lebih sedikit contoh serupa untuk mengaburkan pesakit ini, maka rekod mereka lebih menonjol — dan menonjol itulah yang dikesan serangan keahlian. Kegagalan privasi itu bukan rawak; ia tertumpu pada mereka yang sudah berada di pinggir.

Model lebih besar menjadikannya lebih teruk. Bilangan pesakit yang terdedah kepada serangan hampir sempurna meningkat tajam dengan kapasiti model — dalam satu set data dermatologi, bahagiannya meningkat daripada hampir sifar pada model paling kecil kepada kira-kira **satu dalam sepuluh** pada model terbesar. Ketika model AI perubatan menjadi semakin besar dan berkeupayaan — arah seluruh bidang — risiko khusus ini, amaran para penulis, menjadi lebih serius, bukan kurang.

Ringkasan Rückert terus-terang: “Ini bukan risiko yang boleh diterima. Data kesihatan sangat sensitif.”

Mengapa “selamat secara purata” ialah ujian yang salah

Godaan ialah membaca risiko purata rendah sebagai sijil kesihatan bersih. Pelajaran teras kertas ialah bahawa membuat purata di sini bukan sekadar kurang tepat — ia mengukur perkara yang salah.

Jaminan privasi hanya bermakna jika ia terpakai kepada orang yang paling berisiko, bukan orang di tengah. Model apabila 999 daripada 1,000 pesakit tidak boleh dikenal pasti tetapi seorang boleh dikenal hampir sempurna bukan “99.9% peribadi” dalam apa-apa erti yang bermakna bagi orang itu — dan jika orang itu boleh diramal sebagai pesakit penyakit jarang atau pesakit etnik minoriti, metrik tersebut bukan sekadar tidak lengkap, tetapi secara senyap diskriminatif. Ia melaporkan keselamatan bagi majoriti lalu memanggilnya keselamatan untuk semua.

Itulah perubahan yang dipaksa kertas ini: daripada *sejauh mana model ini peribadi secara purata?* kepada *siapakah pesakit yang paling terdedah, dan siapa mereka?* Itu dua soalan berbeza, dan hanya yang kedua ialah soalan privasi sebenar.

Apa yang *tidak* dibuktikan — atau didakwa — oleh kajian ini

- Ia **tidak** mengatakan AI perubatan harus ditinggalkan. Bingkai para penulis ialah mitigasi, bukan pengunduran: ukur dan baiki risiko, jangan berhenti membina.
- Ia **tidak** bermakna setiap model perubatan membocorkan data, atau bahawa mana-mana model yang digunakan sekarang pernah diserang. Ia menunjukkan *risiko itu wujud dan diagihkan secara tidak sama,* dan metrik agregat biasa gagal melihatnya.
- Ia **tidak** bermakna rekod anda sudah terdedah. Serangan memerlukan keadaan khusus — akses kepada model, rekod calon, dan infrastruktur penyerang sendiri — bukan keupayaan kasual.
- Ia **tidak** menunjukkan *kandungan* rekod pesakit bocor. Yang bocor ialah keahlian — fakta bahawa rekod termasuk dalam data — yang berbahaya atas sebab berbeza, bukan kerana rekod dikeluarkan.
- Ia **tidak** mereduksi jurang kepada satu nombor tajuk utama. Hasil kukuh dan boleh diulang ialah *pola* — metrik agregat meremehkan risiko individu, dan pesakit kurang diwakili menanggung bahagian terbesar — merentas set data dan ratusan model.

Sejauh mana kukuh buktinya?

Ini hasil empirikal dan metodologi, dan kukuh. Pola — metrik privasi agregat secara sistematik meremehkan risiko per pesakit, risiko baki tertumpu pada kumpulan kurang diwakili, dan bertambah buruk dengan kapasiti model — kekal merentas **tujuh set data daripada jenis data berbeza** dan sebilangan besar model bagi setiap set. Keluasan itulah yang menjadikan dakwaan pengukuran ini kredibel dan bukan artifak satu set data.

Dua peringatan jujur. Pertama, ini demonstrasi *risiko*, diukur dengan menjalankan serangan yang dibina para penulis sendiri; ia mencirikan sejauh mana baik penyerang berkeupayaan *boleh* berjaya di bawah andaian yang dinyatakan, bukan berapa kerap serangan sebenar berlaku. Kedua, nombor yang dramatik — kejayaan hampir sempurna bagi sesetengah pesakit — menerangkan individu paling terdedah *secara sengaja*; itulah keseluruhan intinya, dan ia patut dibaca sebagai "ekor jauh lebih berat daripada yang disiratkan purata," bukan "kebanyakan pesakit terdedah."

Sikap yang sesuai bukan panik atau penolakan: ini kajian berhati-hati merentas banyak set data yang menunjukkan cara piawai kita mengesahkan privasi AI perubatan buta kepada kes terburuknya sendiri — dan kes terburuk itu jatuh pada pesakit yang mempunyai ruang keselamatan paling kecil.

Mengapa ia penting

Dua perkara menjadikan ini lebih daripada nota kaki teknikal.

Pertama, ia mengubah apa yang patut dibenarkan bermaksud "memelihara privasi." Jika model dilepaskan dengan skor privasi purata, skor itu boleh benar-benar rendah tetapi masih menyembunyikan subset pesakit yang hampir sempurna boleh dikenal pasti oleh serangan keahlian. Tuntutan praktikal kertas itu konkrit: nilai risiko privasi **bagi setiap pesakit** sebelum pelepasan, kawal siapa yang boleh mengakses model yang digunakan, dan gunakan teknik seperti **privasi pembezaan** — bunyi kecil yang ditentukur dengan teliti dan ditambah semasa latihan untuk menumpulkan serangan keahlian, dengan kos sebenar dan terurus kepada kegunaan model (pertukaran privasi–utiliti dinyatakan secara eksplisit, bukan percuma). "Kami memeriksa purata" tidak patut lagi dikira sebagai sudah memeriksa.

Kedua, ia mengikat privasi dengan keadilan. Kumpulan sama yang kurang diwakili dalam data perubatan — dan oleh itu sudah dilayan lebih buruk oleh AI perubatan — rupanya ialah kumpulan yang privasinya paling kurang dilindungi oleh AI tersebut. Bidang yang berusaha keras menangani ketidaksamaan pertama tidak boleh menganggap yang kedua sebagai masalah jabatan lain. Mereka ialah orang yang sama.

Cerita meyakinkan tentang privasi AI perubatan — *kami mengukurnya, puratanya rendah, jadi semuanya baik* — ialah cerita yang diuraikan oleh kertas ini. Bukan untuk menakutkan orang daripada teknologi, tetapi untuk menaikkan piawaian ke tempat sepatutnya: satu janji yang hanya boleh anda dakwa telah ditunaikan jika anda memeriksanya bagi orang yang paling mungkin terluka.

Ringkasan utama

Serangan inferens keahlian cuba menentukan sama ada rekod seseorang tertentu berada dalam data latihan model AI — dan kerana keahlian itu sendiri boleh mendedahkan sesuatu (contohnya bahawa anda mempunyai penyakit tertentu), ia ialah kebocoran privasi sebenar walaupun rekod asas tidak pernah muncul. Kerja terdahulu mengukur berapa kerap serangan seperti itu berjaya *secara purata*

merentas set data. Kajian ini mengukurnya *bagi setiap pesakit*, merentas tujuh set data perubatan (pengimejan, ECG, rekod elektronik) dan banyak model bagi setiap set, dan mendapati metrik agregat sangat meremehkan risiko: sesetengah individu boleh dikenal pasti hampir sempurna (AUC serangan ≥ 0.95) walaupun purata keseluruhan kelihatan selamat; yang paling terdedah secara sistematik datang daripada kumpulan kurang diwakili (etnik minoriti, penyakit jarang, pengimejan luar biasa); dan masalah bertambah dengan saiz model. Para penulis tidak berhujah untuk meninggalkan AI perubatan — mereka berhujah untuk mengukur privasi pada peringkat individu, mengawal akses model, dan menggunakan privasi pembezaan. Kesimpulannya: "peribadi secara purata" bukan jaminan privasi, dan orang yang gagal dilindunginya biasanya ialah mereka yang memang paling kurang dilindungi.

Semakan tanpa gembar-gembur

Apa yang ditunjukkan oleh kertas ini: Merentas tujuh set data perubatan dan banyak model bagi setiap set, risiko inferens keahlian per pesakit jauh lebih tinggi bagi sesetengah individu daripada yang disiratkan metrik agregat — sehingga kejayaan serangan hampir sempurna (AUC ≥ 0.95) — dan risiko baki itu secara tidak seimbang menimpa kumpulan kurang diwakili serta meningkat dengan kapasiti model.

Apa yang munasabah tetapi belum terbukti: Bahawa penyerang dunia sebenar sudah mengeksploitasi perkara ini terhadap model klinikal yang digunakan; kajian menunjukkan *keupayaan dan taburannya* di bawah andaian tertentu, bukan kekerapan serangan di dunia sebenar.

Apa yang tidak ditunjukkannya: Bahawa AI perubatan patut ditinggalkan; bahawa setiap model bocor atau mana-mana model tertentu yang digunakan telah diceroboh; bahawa *kandungan* rekod pesakit (bukan keahlian) terdedah; satu angka jurang tunggal — hasil kukuh ialah pola, bukan satu nombor.

Had utama: Ia mengukur risiko kes terburuk melalui serangan yang dibina para penulis sendiri, bukan insiden yang diperhatikan; angka dramatik menerangkan individu paling terdedah secara sengaja; dan jurang tepat antara kumpulan ditunjukkan sebagai pola konsisten merentas set data, bukan direduksi kepada satu nombor.

Berapa besar keyakinan yang patut ada pada pembaca umum? Tinggi bahawa metrik privasi agregat meremehkan risiko individu, bahawa risiko baki tertumpu pada pesakit kurang diwakili, dan bahawa masalah ini bertambah dengan saiz model — ia ditunjukkan merentas banyak set data dan model. Sederhana tentang kekerapan serangan sebenar, yang tidak diukur kajian ini. Bacaan selamat: bukan "AI perubatan membocorkan data anda," dan bukan "privasi sudah diselesaikan," tetapi "pemeriksaan privasi piawai buta kepada kes terburuknya sendiri, dan kes itu menimpa pesakit paling terdedah — jadi pemeriksaan itu perlu 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