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

Một AI y tế có thể "riêng tư trung bình" mà vẫn phơi lộ từng bệnh nhân — đặc biệt là nhóm thiếu đại diện

Một mô hình AI y tế được gọi là "bảo vệ quyền riêng tư" thường dựa trên một con số trung bình. Nghiên cứu này lập luận đó là phép kiểm tra sai. Với membership inference attack — suy ra hồ sơ một người cụ thể có nằm trong dữ liệu huấn luyện hay không, từ đó có thể tiết lộ họ từng mắc một bệnh — các tác giả đo rủi ro theo từng bệnh nhân thay vì gộp, trên bảy bộ dữ liệu y tế (ảnh, ECG, bệnh án điện tử) và nhiều mô hình mỗi bộ. Mẫu hình: mô hình trông an toàn trung bình vẫn có thể cho kẻ tấn công nhận diện một số cá nhân gần như hoàn hảo (AUC tấn công ≥ 0,95); người bị phơi lộ có hệ thống thuộc nhóm thiếu đại diện (sắc tộc thiểu số, bệnh hiếm, hình ảnh bất thường); và vấn đề tệ hơn khi mô hình lớn. Tác giả không nói bỏ AI y tế — họ nói phải đo quyền riêng tư theo từng bệnh nhân, kiểm soát truy cập mô hình và dùng differential privacy. Cốt lõi khó chịu: "riêng tư trung bình" không phải bảo đảm quyền riêng tư, và nó thất bại chính với những bệnh nhân vốn ít được bảo vệ nhất.

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

Một cam kết về quyền riêng tư là lời hứa với từng người, không phải với một giá trị trung bình

Khi một mô hình AI y tế được gọi là "bảo vệ quyền riêng tư", tuyên bố đó thường dựa trên một con số: xét trên tất cả bệnh nhân có dữ liệu dùng để huấn luyện, xác suất trung bình để suy ra một cá nhân có nằm trong tập huấn luyện hay không là thấp. Nghe khá yên tâm. Điểm yên lặng nhưng khó chịu của bài này là: đó là con số sai để kiểm tra.

Quyền riêng tư không phải giá trị trung bình. Nó là lời hứa dành cho từng cá nhân — rằng việc hồ sơ của họ nằm trong tập huấn luyện sẽ không quay lại phơi lộ họ. Một trung bình có thể giữ lời với gần như tất cả nhưng phá vỡ hoàn toàn với một số ít. Các nhà nghiên cứu đo rủi ro quyền riêng tư từng bệnh nhân, ở độ phân giải trước đây chưa được dùng, và tìm đúng điều đó: mô hình trông an toàn khi gộp chung vẫn có thể làm lộ việc một số cá nhân thuộc tập huấn luyện gần như hoàn hảo — và những người bị lộ không cân xứng là những nhóm vốn đã ít được bảo vệ nhất.

Các tác giả đã làm gì

Nhóm — do Moritz Knolle dẫn đầu, cùng Daniel Rückert (đều tại Technical University of Munich) và Georg Kaissis (Hasso Plattner Institute), với đồng nghiệp tại Imperial College London — lấy một mối đe dọa quyền riêng tư đã biết là **membership inference attack** và đổi câu hỏi. Thay vì hỏi "trung bình trên toàn bộ dữ liệu, tấn công thành công bao nhiêu?", họ hỏi "với *bệnh nhân cụ thể này*, họ bị phơi lộ đến mức nào?"

Họ chạy phân tích theo từng bệnh nhân trên **bảy bộ dữ liệu y khoa đã được sử dụng rộng rãi**, bao gồm các loại dữ liệu rất khác nhau — ảnh y khoa, điện tâm đồ và bệnh án điện tử — với 200 mô hình cho mỗi bộ. Với từng bệnh nhân ở từng mô hình, họ ước tính mức độ chắc chắn mà kẻ tấn công có thể biết hồ sơ người đó từng là một phần dữ liệu huấn luyện, rồi chia kết quả theo nhóm: tình trạng bệnh, chủng tộc tự khai, bảo hiểm, giới tính và protocol chụp ảnh.

> **Membership inference attack thực sự là gì**
>
> Sự rò rỉ ở đây tinh tế hơn "mô hình nhả hồ sơ của bạn". Nó nói về *membership* — chỉ riêng sự thật dữ liệu của bạn từng nằm trong tập huấn luyện.
> Hãy bắt đầu từ việc một mô hình như vậy thường làm. Bạn đưa vào một phim chụp — chẳng hạn X-quang ngực — và nó trả xác suất: 78% viêm phổi, cùng các giá trị cho tim to, phù, đông đặc. Câu trả lời chẩn đoán thông thường đó là *thứ duy nhất* cuộc tấn công sử dụng. Không có gì đặc biệt.
> Điểm yếu mà nó dựa vào là: một mô hình thường **tự tin hơn một chút** với đúng những ví dụ từng dùng để huấn luyện so với ví dụ chưa từng thấy. Hãy tưởng tượng một học sinh bí mật xem đề trước — với câu đã luyện, câu trả lời đến nhanh và chắc hơn một

chút. Suy ra từ dấu hiệu đó xem một hồ sơ cụ thể có nằm trong những ví dụ mô hình từng học hay không chính là "membership inference".

Cuộc tấn công diễn ra thế này. Ai đó muốn biết phim chụp của một người có được dùng để huấn luyện mô hình không. Họ **không** hỏi mô hình trả hồ sơ — họ đã có một ảnh ứng viên (ảnh của chính người đó hoặc bản rất gần). Họ gửi nó vào một lần như yêu cầu chẩn đoán bình thường và ghi lại mức tự tin. Sau đó họ xem mức tự tin đó giống mô hình *đã* huấn luyện trên ảnh hơn hay mô hình *chưa* huấn luyện hơn — so sánh mà họ có thể tạo khá rẻ bằng một "mô hình tham chiếu" tự huấn luyện để học dấu hiệu quá tự tin, trên máy tính thường không cần phần cứng đặc biệt. Nếu mô hình thật chắc bất thường về ảnh này — như thể "nhận ra" nó — đó là bằng chứng mạnh ảnh từng nằm trong dữ liệu huấn luyện.

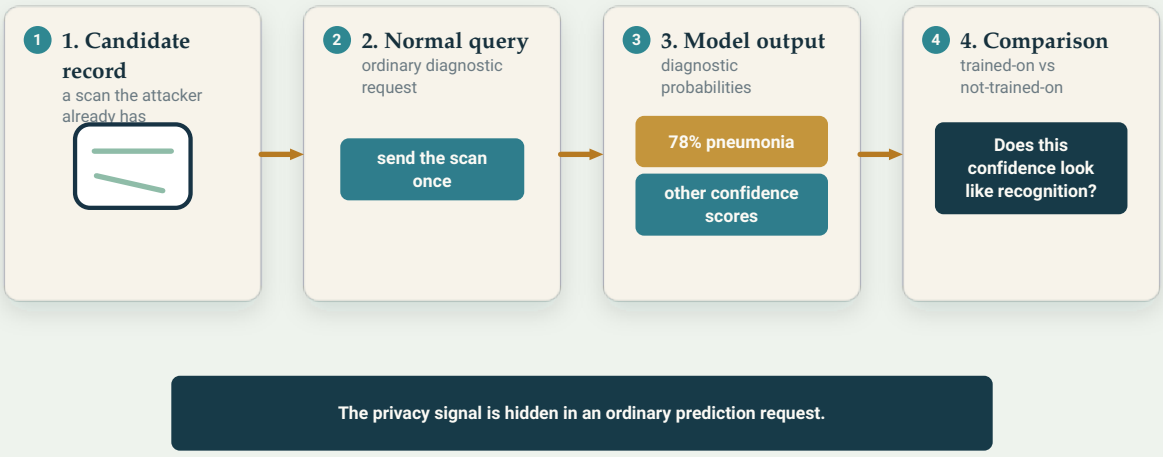
Điều khó chịu là yêu cầu trông không giống tấn công: nó là cùng câu hỏi chẩn đoán mà bác sĩ sẽ gửi, còn tín hiệu quyền riêng tư ẩn trong một dự đoán bình thường. Đó chính là lý do rủi ro cụ thể — dù đến nay nó được chứng minh dưới các giả định phòng thí nghiệm đã nêu, chưa phải vụ tấn công bắt gặp ngoài đời.

Vì sao membership lại quan trọng nếu nội dung hồ sơ không bao giờ rò? Vì *membership bản thân là một sự thật*. Nếu mô hình được huấn luyện trên bệnh nhân nhận một liệu pháp miễn dịch ung thư cụ thể, xác nhận hồ sơ của bạn ở trong đó tiết lộ rằng có lẽ bạn từng mắc loại ung thư ấy — loại thông tin một hãng bảo hiểm hay nhà tuyển dụng không bao giờ nên suy ra. Nội dung vẫn đóng kín; chính sự thuộc về tập dữ liệu là thứ thoát ra.

Các tác giả nêu rõ các giả định — truy cập dự đoán thông thường của mô hình, có hồ sơ ứng viên và mô hình tham chiếu của kẻ tấn công — không phải như hướng dẫn thực hiện, mà để mối đe dọa có thể được phân tích thay vì nói mơ hồ.

A membership attack can hide inside a normal prediction

The attacker does not ask for the record. They already hold a candidate and query the model once.



Mechanism only: no pseudocode, no attack threshold, no recipe.

Tấn công không cần API quyền riêng tư đặc biệt. Một truy vấn chẩn đoán bình thường có thể làm lộ tín hiệu membership nếu mô hình tự tin bất thường với hồ sơ đã từng thấy.

Họ tìm thấy gì

Ba phát hiện, mỗi cái sắc hơn cái trước.

Trung bình che giấu người bị phơi lộ. Khi đo gộp — cách thông thường — nhiều mô hình trông khá yên tâm: với phần lớn bệnh nhân, tấn công không hơn ngẫu nhiên. Nhưng góc nhìn từng người kể câu chuyện khác. Như Knolle nói trong thông báo của nghiên cứu, các đánh giá trước "chỉ đo rủi ro trung bình trên tất cả bệnh nhân. Chúng tôi lần đầu kiểm tra ở cấp từng bệnh nhân — và bức tranh rất khác." Với một số người, tấn công thành công **gần như hoàn hảo** — AUC tấn công **0,95 hoặc cao hơn** (0,5 như tung đồng xu, 1,0 là hoàn hảo) — ngay cả khi trung bình toàn bộ dữ liệu trông không hơn ngẫu nhiên.

Giá trị trung bình có thể nằm gần mức ngẫu nhiên trong khi một đuôi nhỏ các hồ sơ dễ bị nhận diện hơn rất nhiều. Điểm của bài là quyền riêng tư phải được kiểm tra ở cấp bệnh nhân, không chỉ gộp chung.

Mức phơi lộ không bình đẳng — và rơi vào nhóm thiếu đại diện. Những bệnh nhân dễ bị tấn công gần hoàn hảo nhất có hệ thống thuộc các nhóm **thiếu đại diện trong dữ liệu**: sắc tộc thiểu số, phenotype bệnh hiếm, đặc điểm hình ảnh bất thường. Trong bộ bệnh án điện tử,

bệnh nhân da đen xuất hiện **nhiều hơn kỳ vọng 31%** trong nhóm hồ sơ dễ tổn thương nhất; trong bộ mammography, các ảnh bị đánh dấu nghi ngờ ác tính bị đại diện quá mức trong vùng nguy hiểm đó tới **+1.179%**. (Hai con số này là ví dụ, không phải toàn bộ bức tranh — bài báo thấy cùng thiên lệch ở nhiều nhóm — và chúng là mức *quá đại diện tương đối* trong cái đuôi nhỏ rủi ro cực cao: những bệnh nhân này xuất hiện nhiều hơn bao nhiêu trong nhóm bị phơi lộ nhất, **không phải** phần trăm toàn bộ bệnh nhân da đen hay ảnh mammography nghi ngờ bị phơi lộ.) Mô hình có ít ví dụ tương tự để làm những hồ sơ này "chìm" vào đám đông, nên chúng nổi bật — và nổi bật đúng là điều membership attack phát hiện. Thất bại quyền riêng tư không ngẫu nhiên; nó tập trung vào người vốn ở rìa.

Mô hình lớn hơn làm vấn đề tệ hơn. Số bệnh nhân phơi lộ trước tấn công gần hoàn hảo tăng mạnh theo dung lượng mô hình — trong một bộ da liễu, tỷ lệ tăng từ gần như 0 ở mô hình nhỏ nhất lên khoảng **một trên mười** ở mô hình lớn nhất. Khi AI y tế ngày càng lớn và mạnh — hướng cả lĩnh vực đang đi — rủi ro cụ thể này, theo cảnh báo tác giả, tăng chứ không giảm.

Tóm tắt của Rückert rất thẳng: "Đây không phải mức rủi ro chấp nhận được. Dữ liệu sức khỏe cực kỳ nhạy cảm."

Vì sao "an toàn trung bình" là phép kiểm tra sai

Cám dỗ là đọc rủi ro trung bình thấp như giấy chứng nhận an toàn. Bài học cốt lõi của bài là phép lấy trung bình ở đây không chỉ thiếu chính xác — nó đang đo nhầm thứ.

Một cam kết quyền riêng tư chỉ có nghĩa nếu nó giữ được với người có rủi ro cao nhất, không phải người ở giữa. Mô hình mà 999/1.000 bệnh nhân không thể bị nhận diện nhưng một người bị nhận diện gần như hoàn hảo không thể gọi "99,9% riêng tư" theo bất kỳ nghĩa nào quan trọng với người đó — và nếu người đó có thể dự đoán là bệnh nhân bệnh hiếm hay thành viên sắc tộc thiểu số, metric không chỉ thiếu mà còn âm thầm phân biệt đối xử. Nó báo an toàn cho đa số rồi gọi đó là an toàn cho tất cả.

Đây là chuyển dịch mà bài buộc ta thực hiện: từ *mô hình này riêng tư trung bình đến đâu?* sang *bệnh nhân nào bị phơi lộ nhất, và họ là ai?* Hai câu hỏi khác nhau, và chỉ câu thứ hai thật sự là câu hỏi quyền riêng tư.

Điều nghiên cứu này *không* chứng minh — hoặc tuyên bố

- Nó **không** nói nên bỏ AI y tế. Khung của tác giả là giảm thiểu, không rút lui: đo và sửa rủi ro, không ngừng xây dựng.
- Nó **không** có nghĩa mọi mô hình y tế đều rò, hoặc bất kỳ mô hình đang triển khai nào đã bị tấn công. Nó cho thấy *rủi ro tồn tại và phân bố không đều*, còn metric gộp thông thường

bỏ sót nó.

- Nó **không** có nghĩa hồ sơ của bạn đã bị lộ. Tấn công cần điều kiện cụ thể — truy cập mô hình, có hồ sơ ứng viên và hạ tầng riêng của kẻ tấn công — không phải năng lực ngẫu nhiên.
- Nó **không** cho thấy *nội dung* hồ sơ bệnh nhân rò. Thứ rò là membership — sự thật về việc được đưa vào — nguy hiểm vì lý do khác, không phải hồ sơ bị "đổ" ra.
- Nó **không** rút chênh lệch về một con số tiêu đề duy nhất. Kết quả mạnh và lặp lại được là *mẫu hình* — metric gộp đánh giá thấp rủi ro từng người, và bệnh nhân thiếu đại diện gánh nhiều nhất — trên nhiều bộ dữ liệu và hàng trăm mô hình.

Bằng chứng mạnh đến đâu?

Đây là kết quả thực nghiệm và phương pháp luận, và khá vững. Mẫu hình — metric gộp đánh giá thấp có hệ thống rủi ro từng bệnh nhân, phần rủi ro còn lại tập trung ở nhóm thiếu đại diện và tệ hơn khi mô hình lớn — giữ được trên **bảy bộ dữ liệu thuộc các loại khác nhau** và số lượng lớn mô hình cho mỗi bộ. Chính độ rộng này khiến một tuyên bố về đo lường đáng tin hơn một artefact của một bộ dữ liệu.

Có hai lưu ý trung thực. Thứ nhất, đây là chứng minh *rủi ro*, đo bằng cách chạy chính các tấn công tác giả xây; nó mô tả một kẻ tấn công đủ năng lực *có thể* làm tốt đến đâu dưới giả định đã nêu, không đo tấn công thực tế xảy ra thường xuyên đến mức nào. Thứ hai, các con số nổi bật — tấn công gần hoàn hảo với một số bệnh nhân — theo thiết kế mô tả những người tệ nhất; đó chính là điểm của bài, và nên đọc là "đuôi rủi ro nặng hơn trung bình rất nhiều", không phải "phần lớn bệnh nhân bị lộ".

Lập trường đúng không phải hoảng sợ hay gạt bỏ: đây là nghiên cứu cẩn thận trên nhiều bộ dữ liệu cho thấy cách chuẩn chúng ta chứng nhận quyền riêng tư AI y tế bị mù trước những ca xấu nhất của chính nó — và ca xấu nhất rơi vào những bệnh nhân có ít dư địa an toàn nhất.

Vì sao điều này quan trọng

Hai điều khiến đây không chỉ là chú thích kỹ thuật.

Thứ nhất, nó thay đổi điều "bảo vệ quyền riêng tư" nên được phép có nghĩa. Nếu mô hình được phát hành với điểm quyền riêng tư trung bình, điểm đó có thể thật sự thấp mà vẫn che một nhóm bệnh nhân gần như bị nhận diện hoàn hảo bởi membership attack. Yêu cầu thực tế của bài rất cụ thể: đánh giá rủi ro quyền riêng tư **theo từng bệnh nhân** trước phát hành, kiểm soát ai được truy cập mô hình triển khai và dùng kỹ thuật như **differential privacy** — thêm nhiễu nhỏ, hiệu chỉnh cẩn thận trong huấn luyện để làm cùn membership attack, với chi phí thật và có quản lý lên độ hữu dụng của mô hình (trade-off quyền riêng tư–tiện ích là rõ ràng, không miễn phí). "Chúng tôi đã kiểm tra trung bình" không nên còn được tính là đã kiểm tra.

Thứ hai, nó bện quyền riêng tư với công bằng. Cùng những nhóm thiếu đại diện trong dữ liệu

y tế — và vì vậy vốn đã được AI y tế phục vụ kém hơn — hóa ra là nhóm có quyền riêng tư được AI bảo vệ kém nhất. Một lĩnh vực đang cố sửa bất bình đẳng thứ nhất không thể coi bất bình đẳng thứ hai là việc của phòng ban khác. Đó là cùng những con người.

Câu chuyện dễ chịu về quyền riêng tư AI y tế — *chúng ta đã đo, trung bình thấp, vậy ổn* — chính là câu chuyện bài này tháo ra. Không phải để dọa mọi người rời công nghệ, mà để chuyển tiêu chuẩn đến đúng chỗ: một lời hứa chỉ có thể tuyên bố đã giữ khi đã kiểm tra với người có khả năng bị tổn thương nhất.

Tóm tắt gọn

Membership inference attack cố xác định hồ sơ của một người cụ thể có nằm trong dữ liệu huấn luyện của mô hình AI hay không — và vì bản thân membership có thể tiết lộ thông tin (chẳng hạn bạn từng mắc một bệnh cụ thể), đây là rò rỉ quyền riêng tư thật ngay cả khi hồ sơ nền không bao giờ xuất hiện. Công trình trước đo tấn công thành công *trung bình* trên dữ liệu. Nghiên cứu này đo *theo từng bệnh nhân* trên bảy bộ dữ liệu y tế (ảnh, ECG, bệnh án điện tử) và nhiều mô hình mỗi bộ, rồi thấy metric gộp đánh giá thấp rủi ro nghiêm trọng: một số cá nhân có thể bị nhận diện gần như hoàn hảo (AUC tấn công $\geq$ 0,95) dù trung bình toàn bộ trông an toàn; người dễ tổn thương nhất có hệ thống thuộc nhóm thiếu đại diện (sắc tộc thiểu số, bệnh hiếm, hình ảnh bất thường); và vấn đề tăng khi mô hình lớn. Tác giả không kêu gọi bỏ AI y tế — họ kêu gọi đo quyền riêng tư ở cấp cá nhân, kiểm soát truy cập mô hình và dùng differential privacy. Điều cần nhớ: "riêng tư trung bình" không phải bảo đảm quyền riêng tư, và những người nó thất bại thường chính là nhóm vốn ít được bảo vệ nhất.

Đánh giá thẳng thắn

Bài báo cho thấy gì: Trên bảy bộ dữ liệu y tế và nhiều mô hình mỗi bộ, rủi ro membership inference theo bệnh nhân cao hơn rất nhiều với một số cá nhân so với metric gộp — đến mức tấn công gần hoàn hảo (AUC $\geq$ 0,95) — và phần rủi ro dư rơi không cân xứng vào nhóm thiếu đại diện, tăng theo dung lượng mô hình.

Điều hợp lý nhưng chưa được chứng minh: Kẻ tấn công ngoài đời đã khai thác điều này với các mô hình lâm sàng đang triển khai; nghiên cứu chứng minh *năng lực và phân bố của nó* dưới giả định nêu rõ, không đo tần suất tấn công thực tế.

Điều nghiên cứu không cho thấy: Nên bỏ AI y tế; mọi mô hình đều rò hoặc một mô hình triển khai cụ thể đã bị xâm phạm; *nội dung* hồ sơ bệnh nhân (thay vì membership) bị phơi lộ; một con số chênh lệch duy nhất — kết quả vững là mẫu hình, không phải một số.

Hạn chế chính: Đo rủi ro xấu nhất qua các tấn công do chính tác giả triển khai, không phải sự cố quan sát ngoài đời; các số ấn tượng theo thiết kế mô tả người bị phơi lộ nhất; và chênh lệch chính xác theo nhóm được thể hiện như mẫu hình nhất quán trên bộ dữ liệu, không nén

thành một số duy nhất.

Độc giả phổ thông nên tin ở mức nào? Cao rằng metric quyền riêng tư gộp đánh giá thấp rủi ro cá nhân, phần rủi ro còn lại tập trung ở bệnh nhân thiếu đại diện và tệ hơn khi mô hình lớn — điều này được chứng minh trên nhiều dữ liệu và mô hình. Trung bình về tần suất tấn công ngoài đời, vì nghiên cứu không đo điều đó. Cách đọc an toàn: không phải "AI y tế làm rò dữ liệu của bạn", cũng không phải "quyền riêng tư đã được giải quyết", mà là "phép kiểm tra quyền riêng tư chuẩn bị mù trước chính các ca xấu nhất, và các ca ấy rơi vào bệnh nhân dễ tổn thương nhất — nên phép kiểm tra phải thay đổi."

Nguồn

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