The Clean Paper

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

Uma IA médica pode ser privada em média e ainda expor pacientes específicos, sobretudo os sub-representados

Um modelo de IA médica chamado preservador de privacidade costuma se apoiar em um número médio. Este estudo argumenta que esse é o teste errado. Ao estudar ataques de inferência de participação — que revelam se o registro de uma pessoa específica esteve nos dados de treinamento de um modelo e portanto podem denunciar que ela teve certa doença —, os autores mediram o risco por paciente, não em agregado, em sete conjuntos de dados médicos e muitos modelos. O padrão: modelos que parecem seguros em média ainda podem permitir identificar indivíduos específicos quase perfeitamente (AUC de ataque >= 0,95); os expostos pertencem sistematicamente a grupos sub-representados; e o problema piora à medida que os modelos crescem. Eles não dizem para abandonar a IA médica: dizem para medir a privacidade por paciente, controlar o acesso ao modelo e usar privacidade diferencial.

Written by Lucio Vaglio · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso · 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

Uma garantia de privacidade é uma promessa a uma pessoa, não a uma média

Quando um modelo de IA médica é chamado de "preservador de privacidade", essa afirmação geralmente se apoia em um número: entre todos os pacientes cujos dados treinaram o modelo, a chance média de inferir a participação de qualquer indivíduo é baixa. Isso soa tranquilizador. O ponto discreto e desconfortável deste artigo é que esse é o número errado.

Privacidade não é uma média. É uma promessa feita a cada indivíduo — a de que estar no conjunto de treinamento não voltará para expô-lo. E uma média pode cumprir essa promessa para quase todos enquanto a quebra completamente para alguns. Os pesquisadores mediram o risco de privacidade paciente por paciente, em uma resolução que ninguém havia usado antes, e encontraram exatamente isso: modelos que parecem seguros no agregado podem vazar a participação de indivíduos específicos quase perfeitamente — e os indivíduos que vazam são, desproporcionalmente, aqueles que já eram os menos protegidos.

O que os autores fizeram

A equipe — liderada por Moritz Knolle, com Daniel Rückert (ambos da Universidade Técnica de Munique) e Georg Kaissis (Instituto Hasso Plattner), ao lado de colegas do Imperial College London — pegou uma ameaça de privacidade conhecida, chamada **ataque de inferência de participação**, e mudou a pergunta. Em vez de perguntar "em média, com que frequência esse ataque tem sucesso no conjunto de dados?", perguntou "para *este paciente específico*, quão exposto ele está?"

Eles executaram essa análise por paciente em **sete bases médicas estabelecidas**, cobrindo tipos de dados muito diferentes — imagens médicas, eletrocardiogramas e prontuários eletrônicos — e, em cada uma delas, 200 modelos. Para cada paciente em cada modelo, estimaram quão confiantemente um atacante poderia dizer se o registro daquela pessoa havia feito parte dos dados de treinamento, depois separaram os resultados por grupo: status de doença, raça autodeclarada, seguro de saúde, sexo e protocolo de imagem.

> **O que é, de fato, um ataque de inferência de participação**
>
> O vazamento aqui é mais sutil do que "o modelo cospe seu prontuário". Trata-se de *participação* — o simples fato de seus dados estarem no conjunto de treinamento.
>
> Comece pelo que um desses modelos normalmente faz. Você entrega uma imagem — uma radiografia de tórax, digamos — e ele devolve probabilidades: 78% de chance de pneumonia, junto com leituras para cardiomegalia, edema, consolidação. Essa resposta diagnóstica cotidiana é a *única* coisa que o ataque usa. Nada exótico.
>
> A fraqueza em que ele se apoia é esta: um modelo costuma ficar um pouco **mais confiante** nos exemplos exatos em que foi treinado do que em exemplos que nunca viu. Pense em um aluno que secretamente viu a prova antes — nas perguntas que praticou, as respostas vêm um pouco

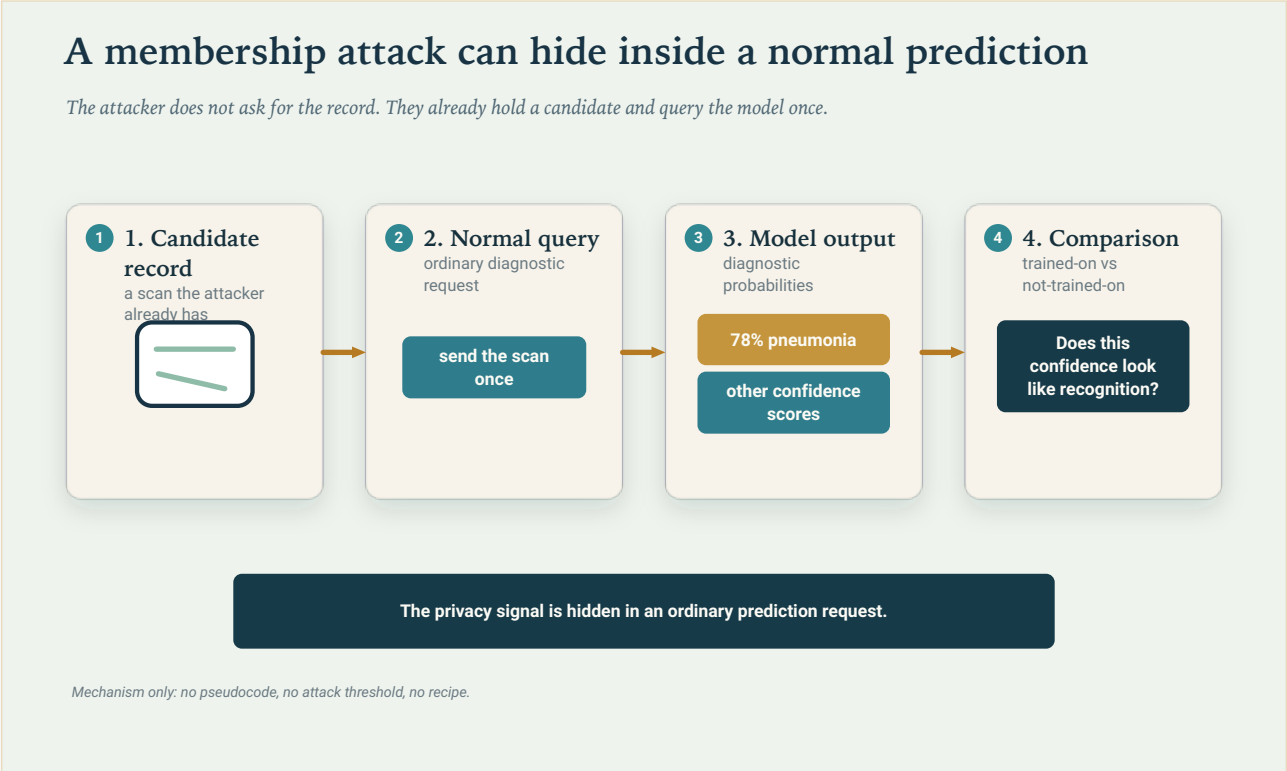
rápido demais, um pouco confiantes demais. Descobrir, a partir desse indício, se um registro específico estava entre os exemplos que o modelo estudou é o que "inferência de participação" significa.

Então aqui está o ataque, passo a passo. Alguém quer saber se a imagem de uma pessoa específica foi usada para treinar o modelo. Ele **não** pede o registro ao modelo — já tem uma imagem candidata (a da própria pessoa, ou uma cópia próxima). Envia-a uma vez, como uma solicitação diagnóstica comum, e observa quão confiante é a resposta. Depois verifica se essa confiança se parece mais com a de um modelo que *foi* treinado com a imagem ou com a de um que *não foi* — uma comparação que pode fazer de forma barata treinando um substituto próprio (um "modelo de referência") para aprender como se parece esse excesso de confiança revelador, em um computador comum sem hardware especial. Se o modelo real fica incomumente seguro sobre essa imagem — como se a reconhecesse —, isso é forte evidência de que a imagem estava nos dados de treinamento.

A parte inquietante é que nada na solicitação parece um ataque: é a mesma consulta diagnóstica que um clínico faria, e o sinal de privacidade está escondido dentro de uma predição comum. É exatamente por isso que o risco é concreto — embora, até aqui, tenha sido demonstrado sob premissas laboratoriais declaradas, não flagrado no mundo real.

Por que a participação importa, se o registro em si nunca vaza? Porque *participação é um fato*. Se um modelo foi treinado com pacientes que receberam uma imunoterapia específica contra câncer, confirmar que seu registro está nele revela que você provavelmente teve aquele câncer — o tipo de coisa que seguradora ou empregador jamais deveria conseguir inferir. O conteúdo fica fechado; o fato de pertencer é o que escapa.

Os autores apresentam essas premissas com clareza — acesso às predições comuns do modelo, um registro candidato e o modelo de referência do próprio atacante — não como um manual, mas para que a ameaça possa ser analisada em vez de descartada com um gesto.



O ataque não precisa de uma API especial de privacidade. Uma consulta diagnóstica normal pode vazar um sinal de participação se o modelo estiver incomumente confiante em um registro que já viu.

O que eles encontraram

Três achados, cada um mais afiado que o anterior.

Médias escondem os expostos. Medidos em agregado — o modo usual — muitos desses modelos parecem tranquilizadoramente privados: o ataque não vai melhor que o acaso para a maioria dos pacientes. A visão por paciente contou outra história. Como Knolle disse no anúncio do estudo, avaliações anteriores "sempre mediram apenas o risco médio em todos os pacientes. Examinamos o risco no nível de pacientes individuais pela primeira vez — e isso pinta um quadro muito diferente." Para alguns indivíduos, o ataque tem sucesso **quase perfeito** — os autores medem isso como uma AUC de ataque de **0,95 ou maior** (0,5 é cara ou coroa, 1,0 é impecável) — mesmo quando a média do conjunto de dados parecia não passar do acaso.

A média pode ficar perto do acaso enquanto uma pequena cauda de registros permanece muito mais exposta. O ponto do artigo é que a privacidade precisa ser checada no nível do paciente, não apenas em agregado.

Original Aurora diagram — The Clean Paper CC BY 4.0

A exposição é desigual — e recai sobre os sub-representados. Os pacientes mais vulneráveis a um ataque quase perfeito eram sistematicamente aqueles de grupos **sub-representados nos dados**: minorias étnicas, fenótipos de doenças raras, características incomuns de imagem. No conjunto de prontuários eletrônicos, pacientes negros apareceram **31% mais frequentemente** que o esperado entre os registros mais vulneráveis; no conjunto de mamografia, exames sinalizados como suspeitos para malignidade ficaram super-representados nessa zona de perigo por impressionantes **+1.179%**. (Esses dois números são exemplos, não o quadro inteiro — o artigo relata o mesmo viés em vários grupos — e são *super-representações relativas* dentro da pequena cauda de risco extremo: quanto mais frequentemente esses pacientes aparecem entre os registros mais expostos, não a fração de pacientes negros ou de mamografias suspeitas que está exposta.) Um modelo tem menos exemplos parecidos para diluir esses pacientes, então seus registros se destacam — e destacar-se é exatamente o que um ataque de participação detecta. A falha de privacidade não é aleatória; concentra-se nas pessoas que já estão nas margens.

Modelos maiores pioram o problema. O número de pacientes expostos a ataques quase perfeitos subiu acentuadamente com a capacidade do modelo — em uma base de dermatologia, a parcela subiu de essencialmente zero no menor modelo para cerca de **um em dez** no maior. À medida que modelos de IA médica ficam maiores e mais capazes — a direção em que o campo inteiro se move —, esse risco específico, alertam os autores, fica mais grave, não menos.

O resumo de Rückert é direto: “Este não é um risco tolerável. Dados de saúde são altamente sensíveis.”

Por que “seguro em média” é o teste errado

A tentação é ler um risco médio baixo como atestado de saúde. A lição central do artigo é que a média aqui não é apenas imprecisa — ela mede a coisa errada.

Uma garantia de privacidade só é significativa se vale para a pessoa em maior risco, não para a pessoa no meio. Um modelo em que 999 de 1.000 pacientes são irrecuperáveis, mas um pode ser identificado quase perfeitamente, não é “99,9% privado” em nenhum sentido que importe para essa pessoa — e, se essa pessoa é previsivelmente o paciente de doença rara ou de minoria étnica, a métrica não é só incompleta, é discretamente discriminatória. Ela relata segurança para a maioria e chama isso de segurança para todos.

Essa é a mudança que o artigo força: de *quão privado é este modelo em média?* para *quem é o paciente mais exposto, e quem é essa pessoa?* São perguntas diferentes, e só a segunda é uma pergunta de privacidade.

O que isso *não* prova — nem afirma

- Não diz que a IA médica deve ser abandonada. O enquadramento dos autores é mitigação, não recuo: medir e corrigir o risco, não parar de construir.
- Não significa que todo modelo médico esteja vazando, nem que qualquer modelo implantado específico tenha sido atacado. Mostra que o *risco existe e é distribuído de forma desigual*, e que métricas agregadas padrão não o enxergam.
- Não significa que seus prontuários já estejam expostos. O ataque precisa de condições específicas — acesso ao modelo, um registro candidato e a infraestrutura própria do atacante —, não é uma capacidade casual.
- Não mostra que o *conteúdo* do paciente vaza. O que vaza é a participação — o fato de inclusão —, perigosa por outra razão, não porque o prontuário seja despejado.
- Não reduz as disparidades a um único número de manchete. O resultado forte e reproduzível é o *padrão* — métricas agregadas subestimam o risco individual, e pacientes sub-representados carregam a maior parte dele — em vários conjuntos de dados e centenas de modelos.

Quão forte é a evidência?

Este é um resultado empírico e metodológico, e robusto. O padrão — métricas agregadas de privacidade subestimam sistematicamente o risco por paciente, o risco residual se concentra em grupos sub-representados e piora com a capacidade do modelo — apareceu em **sete conjuntos de dados de tipos diferentes** e em um grande número de modelos por conjunto. Essa amplitude é exatamente o que torna uma afirmação de medição crível, em vez de um artefato de uma única base.

Duas ressalvas honestas. Primeiro, é uma demonstração de *risco*, medida executando os ataques que os próprios autores construíram; caracteriza quão bem um atacante capaz *poderia* se sair sob pre-

missas declaradas, não com que frequência ataques reais acontecem. Segundo, os números vívidos — sucesso quase perfeito de ataque para alguns pacientes — descrevem os indivíduos em pior situação *por desenho*; esse é o ponto inteiro, e devem ser lidos como "a cauda é muito mais pesada do que a média sugere", não como "a maioria dos pacientes está exposta".

A postura apropriada não é alarme nem descarte: é um estudo cuidadoso, em múltiplas bases, mostrando que a forma padrão de certificar privacidade em IA médica é cega aos próprios piores casos — e que esses piores casos recaem sobre os pacientes com menos margem de sobra.

Por que importa

Duas coisas tornam isso mais que uma nota técnica.

Primeiro, muda o que "preservador de privacidade" deveria poder significar. Se um modelo é liberado com uma pontuação média de privacidade, essa pontuação pode ser genuinamente baixa e ainda esconder um subconjunto de pacientes quase perfeitamente identificáveis por um ataque de participação. A demanda prática do artigo é concreta: avaliar o risco de privacidade **por paciente** antes da liberação, controlar quem pode acessar modelos implantados e usar técnicas como **privacidade diferencial** — ruído pequeno e cuidadosamente calibrado adicionado durante o treinamento que enfraquece ataques de participação, com um custo real e administrado para a utilidade do modelo (a troca privacidade-utilidade é explícita, não gratuita). "Checamos a média" deveria deixar de contar como ter checado.

Segundo, entrelaça privacidade e justiça. Os mesmos grupos que são sub-representados em dados médicos — e portanto já são pior atendidos por IA médica — acabam sendo aqueles cuja privacidade essa IA menos protege. Um campo que trabalha duro na primeira desigualdade não pode tratar a segunda como departamento de outra pessoa. São as mesmas pessoas.

A história tranquilizadora da privacidade em IA médica — *medimos, a média é baixa, estamos bem* — é a história que este artigo desmonta. Não para afastar ninguém da tecnologia, mas para deslocar o padrão para onde ele pertence: uma promessa que você só pode afirmar cumprir se a verificou para a pessoa mais provável de ser prejudicada.

Resumo limpo

Ataques de inferência de participação tentam determinar se o registro de uma pessoa específica estava nos dados de treinamento de um modelo de IA — e, como a participação pode ser reveladora por si só (por exemplo, que você teve uma doença específica), isso é um vazamento real de privacidade mesmo quando o registro subjacente nunca aparece. Trabalhos anteriores mediam a frequência com que esses ataques têm sucesso *em média* no conjunto de dados. Este estudo mediu isso *por paciente*, em sete conjuntos de dados médicos (imagem, ECG, prontuários eletrônicos) e muitos modelos em cada um, e descobriu que métricas agregadas subestimam muito o risco: alguns indivíduos podem

ser identificados quase perfeitamente (AUC de ataque $\geq 0{,}95$) mesmo quando a média do conjunto parece segura; os mais vulneráveis são sistematicamente aqueles de grupos sub-representados (minoria étnica, doença rara, imagem incomum); e o problema cresce com o tamanho do modelo. Os autores não defendem abandonar a IA médica — defendem medir privacidade no nível individual, controlar o acesso ao modelo e usar privacidade diferencial. A mensagem: "privado em média" não é uma garantia de privacidade, e as pessoas em que falha costumam ser as já menos protegidas.

Checagem sem rodeios

O que o artigo mostra: Em sete conjuntos de dados médicos e muitos modelos em cada um, o risco de inferência de participação por paciente é muito maior para alguns indivíduos do que métricas agregadas sugerem — até sucesso de ataque quase perfeito (AUC $\geq 0{,}95$) — e esse risco residual recai desproporcionalmente sobre grupos sub-representados e cresce com a capacidade do modelo.

O que é plausível, mas não provado: Que atacantes reais já estejam explorando isso contra modelos clínicos implantados; o estudo demonstra a *capacidade e sua distribuição* sob premissas declaradas, não a frequência de ataques no mundo real.

O que não mostra: Que a IA médica deve ser abandonada; que todo modelo vaza ou que algum modelo implantado específico foi violado; que *conteúdos* de prontuários de pacientes (em vez da participação) são expostos; um único número quantificado de disparidade — o resultado robusto é o padrão, não um número.

Principais limitações: Mede risco de pior caso por meio dos ataques dos próprios autores, não incidentes observados; os números dramáticos descrevem os indivíduos mais expostos por desenho; e as disparidades exatas por grupo são mostradas como um padrão consistente entre conjuntos de dados, não reduzidas a um número único.

Quanta confiança um leitor geral deve ter? Alta de que métricas agregadas de privacidade subestimam o risco individual, que o risco residual se concentra em pacientes sub-representados e que isso piora com o tamanho do modelo — demonstrado em muitos conjuntos de dados e modelos. Moderada sobre a frequência real desses ataques, que o estudo não mede. A leitura segura: não "IA médica vaza seus dados" e não "privacidade está resolvida", mas "a checagem padrão de privacidade é cega aos próprios piores casos, e esses casos recaem sobre os pacientes mais vulneráveis — então a checagem precisa mudar."

Fontes

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-ai

Medical Xpress (2026) — coverage of the study
https://medicalxpress.com/news/2026-06-patient-groups-vulnerable-privacy-medical.html