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

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

Chamar «preservador de privacidade» a um modelo de IA médica costuma assentar num único valor médio. Este estudo argumenta que esse é o teste errado. Ao estudar ataques de inferência de pertença — que revelam se o registo de uma pessoa específica esteve nos dados de treino de um modelo e podem, por isso, denunciar que teve determinada doença — os autores mediram o risco por doente em vez de no agregado, em sete conjuntos de dados médicos (imagiologia, ECG e registos eletrónicos) e muitos modelos em cada um. O padrão: modelos que parecem seguros em média podem ainda permitir a identificação quase perfeita de indivíduos específicos (AUC de ataque ≥ 0,95); os expostos pertencem sistematicamente a grupos sub-representados (minorias étnicas, doenças raras, imagiologia invulgar); e o problema agrava-se à medida que os modelos crescem. Os autores não dizem para abandonar a IA médica — dizem para medir a privacidade por doente, controlar o acesso ao modelo e usar privacidade diferencial. O núcleo desconfortável: «privado em média» não é uma garantia de privacidade e falha os doentes que já estão menos protegidos.

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

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

Quando se diz que um modelo de IA médica «preserva a privacidade», essa afirmação costuma assentar num único número: considerando todos os doentes cujos dados treinaram o modelo, a probabilidade média de inferir a pertença de qualquer indivíduo ao conjunto de treino é baixa. Parece 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 pessoa — a promessa de que ter estado no conjunto de treino não voltará mais tarde para a expor. E uma média pode cumprir essa promessa para quase todos enquanto a quebra completamente para alguns. Os investigadores mediram o risco de privacidade doente a doente, com uma resolução que ninguém tinha usado antes, e encontraram exatamente isso: modelos que parecem seguros no agregado podem revelar quase perfeitamente a pertença de indivíduos específicos ao treino — e as pessoas que expõem são, de forma desproporcionada, precisamente as que já estavam menos protegidas.

O que fizeram os autores

A equipa — liderada por Moritz Knolle, com Daniel Rückert (ambos da Universidade Técnica de Munique) e Georg Kaissis (Hasso Plattner Institute), juntamente com colegas do Imperial College London — pegou numa ameaça de privacidade bem conhecida, chamada **ataque de inferência de pertença** (*membership inference attack*), e mudou a pergunta. Em vez de perguntar «em média, com que frequência este ataque tem sucesso no conjunto de dados?», perguntou «para *este doente específico*, qual é o grau de exposição?»

Executaram esta análise por doente em **sete conjuntos de dados médicos estabelecidos**, abrangendo tipos de dados muito diferentes — imagiologia médica, eletrocardiogramas e registos clínicos eletrónicos — e, em cada um deles, 200 modelos. Para cada doente em cada modelo, estimaram com que confiança um atacante conseguiria dizer se o registo dessa pessoa tinha feito parte dos dados de treino; depois separaram os resultados por grupo: doença, raça autoidentificada, seguro, sexo e protocolo de imagiologia.

> **O que é realmente um ataque de inferência de pertença**
>
> A fuga aqui é mais subtil do que «o modelo devolve o seu registo». Trata-se da *pertença* — o simples facto de os seus dados terem estado no conjunto de treino.
> Comecemos pelo que um destes modelos normalmente faz. Entregamos-lhe uma imagem — por exemplo, uma radiografia do tórax — e ele devolve probabilidades: 78% de probabilidade de pneumonia, juntamente com avaliações de cardiomegalia, edema, consolidação. Essa resposta diagnóstica banal é a *única* coisa que o ataque utiliza. Nada de exótico.
> A fraqueza em que se apoia é esta: um modelo tende a estar um pouco **mais confiante** nos exemplos exatos com que foi treinado do que nos que nunca viu. Imagine um estudante que

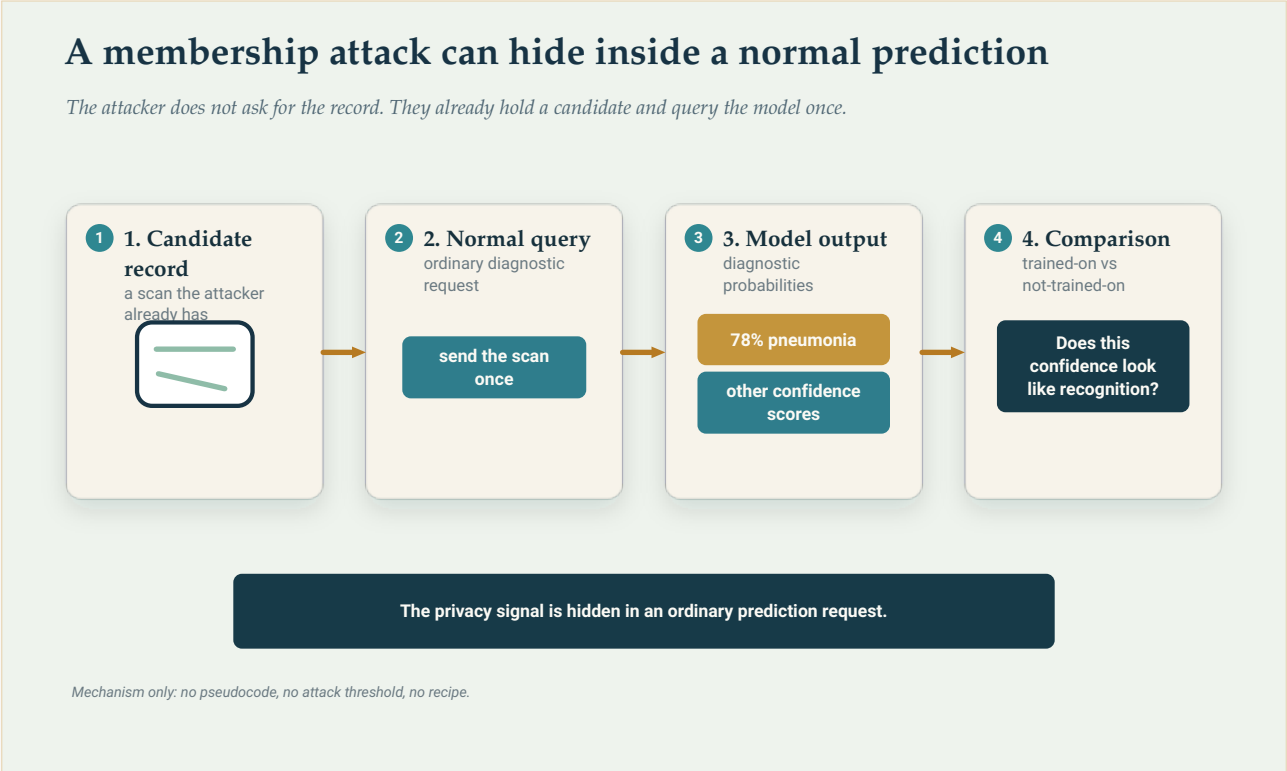
viu secretamente o exame antes da prova — nas perguntas que praticou, as respostas aparecem depressa demais, com confiança a mais. Descobrir, a partir desse indício, se um determinado registo era um dos exemplos que o modelo estudou é aquilo a que se chama «inferência de pertença».

O ataque funciona assim, passo a passo. Alguém quer saber se a imagem de uma determinada pessoa foi usada para treinar o modelo. **Não** pede ao modelo o registo — já possui uma imagem candidata (a da própria pessoa, ou uma cópia muito semelhante). Envia-a uma vez, como um pedido diagnóstico normal, e anota a confiança da 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* — comparação que pode fazer de forma relativamente barata treinando o seu próprio modelo de referência para aprender como se manifesta esse excesso de confiança, num computador normal e sem hardware especial. Se o modelo real estiver invulgarmente seguro acerca daquela imagem — como se a reconhecesse — isso constitui evidência forte de que a imagem esteve nos dados de treino.

A parte inquietante é que nada no pedido parece um ataque: é a mesma consulta diagnóstica que um clínico faria, e o sinal de privacidade está escondido dentro de uma previsão normal. É precisamente isso que torna o risco concreto — embora, até agora, tenha sido demonstrado sob hipóteses laboratoriais explícitas, e não observado em ataques no mundo real.

Porque importa a pertença, se o registo em si nunca sai? Porque *a pertença é um facto*. Se um modelo foi treinado com doentes que receberam uma determinada imunoterapia contra o cancro, confirmar que o seu registo está nele revela que provavelmente teve esse cancro — exatamente o tipo de coisa que uma seguradora ou um empregador nunca deveria conseguir inferir. O conteúdo continua selado; é o facto de pertencer que escapa.

Os autores apresentam estas hipóteses de forma explícita — acesso às previsões normais do modelo, um registo candidato e o próprio modelo de referência do 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 revelar um sinal de pertença se o modelo estiver invulgarmente confiante num registo que já viu durante o treino.

O que encontraram

Três resultados, cada um mais contundente do que o anterior.

As médias escondem quem está exposto. Medidos de forma agregada — a abordagem habitual — muitos destes modelos parecem tranquilizadoramente privados: para a maioria dos doentes, o ataque não funciona melhor do que o acaso. A análise por doente contou outra história. Como Knolle explicou no anúncio do estudo, as avaliações anteriores «sempre mediram apenas o risco médio em todos os doentes. Examinámos pela primeira vez o risco ao nível de cada doente — e o quadro é muito diferente». Para alguns indivíduos, o ataque tem sucesso **quase perfeito** — os autores medem-no como uma AUC de ataque de **0,95 ou superior** (0,5 é equivalente a lançar uma moeda; 1,0 é perfeito) — mesmo quando a média de todo o conjunto de dados não parecia melhor do que o acaso.

A média pode ficar perto do acaso enquanto uma pequena cauda de registos continua muito mais exposta. O ponto do artigo é que a privacidade tem de ser verificada ao nível de cada doente, não apenas no agregado.

Original Aurora diagram — The Clean Paper CC BY 4.0

A exposição é desigual — e atinge os sub-representados. Os doentes mais vulneráveis a um ataque quase perfeito pertenciam sistematicamente a grupos **sub-representados nos dados**: minorias étnicas, fenótipos de doenças raras, características imagiológicas invulgares. No conjunto de dados de registos eletrónicos, doentes negros apareciam **31% mais vezes** do que seria esperado entre os registos mais vulneráveis; no conjunto de mamografias, exames assinalados como suspeitos de malignidade estavam sobrerrepresentados nessa zona de perigo em impressionantes **+1 179%**. (Estes dois valores são exemplos, não o quadro inteiro — o artigo relata o mesmo enviesamento em vários grupos — e são *sobrerrepresentações relativas* dentro da pequena cauda de risco extremo: quantas vezes mais estes doentes aparecem entre os registos mais expostos, e não a percentagem de doentes negros ou de mamografias suspeitas que estão expostos.) Um modelo tem menos exemplos semelhantes em que estes doentes se possam «dissolver», pelo que os seus registos se destacam — e destacar-se é precisamente aquilo que um ataque de pertença deteta. A falha de privacidade não é aleatória; concentra-se nas pessoas que já estão nas margens.

Modelos maiores agravam o problema. O número de doentes expostos a ataques quase perfeitos aumentou fortemente com a capacidade do modelo — num conjunto de dermatologia, a proporção subiu de praticamente zero no modelo mais pequeno para cerca de **um em cada dez** no maior. À medida que os modelos de IA médica se tornam maiores e mais capazes — precisamente a direção em que todo o campo se move — este risco específico, alertam os autores, torna-se mais grave, não menos.

O resumo de Rückert é direto: «Este não é um risco tolerável. Os dados de saúde são altamente

sensíveis.»

Porque «seguro em média» é o teste errado

É tentador ler um risco médio baixo como um certificado de boa saúde. A lição central do artigo é que fazer a média aqui não é apenas impreciso — é medir a coisa errada.

Uma garantia de privacidade só tem significado se também proteger a pessoa mais exposta, não apenas a pessoa mediana. Um modelo em que 999 de 1 000 doentes não podem ser identificados, mas um pode sê-lo quase perfeitamente, não é «99,9% privado» em nenhum sentido que importe para essa pessoa — e, se essa pessoa for previsivelmente o doente com doença rara ou pertencente a uma minoria étnica, a métrica não é apenas incompleta: é silenciosamente discriminatória. Declara segurança para a maioria e chama-lhe segurança para todos.

Essa é a mudança que o artigo força: deixar de perguntar *quão privado é este modelo em média?* e passar a perguntar *quem é o doente mais exposto — e quem é essa pessoa?* São perguntas diferentes, e só a segunda é verdadeiramente uma pergunta sobre privacidade.

O que isto *não* demonstra — nem afirma

- **Não** diz que a IA médica deva ser abandonada. O enquadramento dos autores é mitigação, não retirada: medir e corrigir o risco, não deixar de construir.
- **Não** significa que todos os modelos médicos estejam a expor dados, nem que qualquer modelo já implementado tenha sido atacado. Mostra que o *risco existe e é distribuído de forma desigual*, e que as métricas agregadas habituais não o veem.
- **Não** significa que os seus registos já estejam expostos. O ataque exige condições específicas — acesso ao modelo, um registo candidato e infraestrutura própria do atacante — e não é uma capacidade casual.
- **Não** mostra que o *conteúdo* do registo do doente seja revelado. O que escapa é a pertença — o facto de inclusão — perigoso por uma razão diferente, não porque o registo seja despejado para fora.
- **Não** reduz as disparidades a um único número de manchete. O resultado forte e reproduzível é o *padrão* — as métricas agregadas subestimam o risco individual e os doentes sub-representados suportam a maior parte dele — em vários conjuntos de dados e centenas de modelos.

Quão forte é a evidência?

É um resultado empírico e metodológico, e robusto. O padrão — métricas agregadas de privacidade subestimam sistematicamente o risco por doente, o risco residual concentra-se em grupos sub-representados e piora com a capacidade do modelo — manteve-se em **sete conjuntos de dados de**

tipos diferentes e num grande número de modelos por conjunto. É precisamente essa amplitude que torna uma afirmação de medição credível em vez de um artefacto de uma única base de dados.

Duas ressalvas honestas. Primeiro, é uma demonstração de *risco*, medida executando os ataques construídos pelos próprios autores; caracteriza até que ponto um atacante competente *poderia* ter sucesso sob hipóteses definidas, não com que frequência ataques reais acontecem. Segundo, os números mais impressionantes — sucesso quase perfeito para alguns doentes — descrevem deliberadamente os indivíduos em pior situação; esse é todo o ponto, e devem ser lidos como «a cauda é muito mais pesada do que a média sugere», não como «a maioria dos doentes está exposta».

A postura adequada não é alarme nem desvalorização: é reconhecer um estudo cuidadoso, em vários conjuntos de dados, que mostra que a forma habitual de certificar a privacidade da IA médica é cega aos seus próprios piores casos — e que esses piores casos recaem sobre os doentes com menor margem para suportar o risco.

Porque importa

Duas coisas tornam isto mais do que uma nota técnica.

Primeiro, muda aquilo que «preservar a privacidade» deveria poder significar. Se um modelo é lançado com uma pontuação média de privacidade, essa pontuação pode ser genuinamente baixa e ainda assim esconder um subconjunto de doentes quase perfeitamente identificáveis por um ataque de pertença. A exigência prática do artigo é concreta: avaliar o risco de privacidade **por doente** antes da disponibilização, controlar quem pode aceder aos modelos implementados e usar técnicas como **privacidade diferencial** — uma pequena quantidade de ruído cuidadosamente calibrado acrescentada durante o treino para atenuar ataques de pertença, com um custo real e gerido para a utilidade do modelo (o compromisso privacidade–utilidade é explícito, não gratuito). «Verificámos a média» deveria deixar de contar como verificação suficiente.

Segundo, entrelaça privacidade e equidade. Os mesmos grupos sub-representados nos dados médicos — e, por isso, já pior servidos pela IA médica — são também aqueles cuja privacidade essa IA menos protege. Um campo que trabalha arduamente sobre a primeira desigualdade não pode tratar a segunda como problema de outro departamento. São as mesmas pessoas.

A história tranquilizadora sobre privacidade em IA médica — *medimos, a média é baixa, está tudo bem* — é a história que este artigo desmonta. Não para afastar ninguém da tecnologia, mas para mover o padrão para onde deve estar: uma promessa que só se pode dizer cumprida depois de a verificar para a pessoa com maior probabilidade de ser prejudicada.

Resumo claro

Ataques de inferência de pertença tentam determinar se o registo de uma pessoa específica esteve nos dados de treino de um modelo de IA — e, como essa pertença pode por si só revelar informação

(por exemplo, que a pessoa tinha determinada doença), é uma verdadeira fuga de privacidade mesmo quando o registo subjacente nunca aparece. Trabalhos anteriores mediam com que frequência esses ataques tinham sucesso *em média* num conjunto de dados. Este estudo mediu o risco *por doente*, em sete conjuntos de dados médicos (imagiologia, ECG e registos eletrónicos) e muitos modelos em cada um, e descobriu que as métricas agregadas subestimam fortemente 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 pertencem sistematicamente a grupos sub-representados (minorias étnicas, doenças raras, imagiologia invulgar); e o problema cresce com o tamanho do modelo. Os autores não defendem abandonar a IA médica — defendem medir a privacidade ao nível individual, controlar o acesso aos modelos e utilizar privacidade diferencial. A conclusão: «privado em média» não é uma garantia de privacidade, e falha precisamente as pessoas que já estão menos protegidas.

Verificação sem exageros

O que o artigo mostra: Em sete conjuntos de dados médicos e muitos modelos em cada um, o risco de inferência de pertença por doente é muito superior para alguns indivíduos ao sugerido pelas métricas agregadas — chegando a sucesso quase perfeito do ataque (AUC $\geq 0{,}95$) — e esse risco residual recai de forma desproporcionada sobre grupos sub-representados e aumenta com a capacidade do modelo.

O que é plausível, mas não está demonstrado: Que atacantes no mundo real já estejam a explorar esta vulnerabilidade contra modelos clínicos implementados; o estudo demonstra a *capacidade e a sua distribuição* sob hipóteses explícitas, não a frequência de ataques reais.

O que não mostra: Que a IA médica deva ser abandonada; que todos os modelos tenham fugas ou que um modelo específico em produção tenha sido comprometido; que o *conteúdo* dos registos dos doentes — e não a pertença — esteja exposto; um único número quantificado para a disparidade — o resultado robusto é o padrão, não um valor isolado.

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

Quanta confiança deve ter um leitor geral? Elevada de que as métricas agregadas de privacidade subestimam o risco individual, de que o risco residual se concentra em doentes sub-representados e de que piora com o tamanho do modelo — o padrão é demonstrado em muitos conjuntos de dados e modelos. Moderada quanto à frequência destes ataques no mundo real, que o estudo não mede. A leitura segura não é «a IA médica expõe os seus dados» nem «a privacidade está resolvida», mas «o controlo padrão de privacidade é cego aos próprios piores casos, e esses casos recaem sobre os doentes mais vulneráveis — portanto, o controlo tem de 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 Xpress (2026) — coverage of the study
https://medicalxpress.com/news/2026-06-patient-groups-vulnerable-privacy-medical.html