The CLEAN PAPER

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

医療 AI は平均ではプライベートでも、特定患者を露出させ得る —— とりわけ過小代表の患者を

「プライバシー保護型」と呼ばれる医療 *AI* モデルは通常、一つの平均値で評価される。この研究は、その数字こそ間違った検査だと論じる。特定個人の記録が訓練データに含まれていたかを推測し、その人が特定疾患だったことまで漏らし得るメンバーシップ推論攻撃について、著者らは 7 つの医療データセット（画像、*ECG*、電子記録）と多数のモデルで、集計ではなく患者単位のリスクを測った。平均では安全に見えるモデルでも、特定個人をほぼ完璧に識別できる場合があり（攻撃 $AUC \geq 0.95$）、露出する人は少数民族、希少疾患、珍しい画像特徴などの過小代表集団に体系的に偏り、モデルが大きいほど悪化した。著者らは医療 *AI* を捨てろとは言わない。患者単位でプライバシーを測り、モデルアクセスを制御し、差分プライバシーを使うよう求めている。居心地の悪い核心は、「平均ではプライベート」は保証ではなく、その失敗がすでに最も保護の薄い患者に集中するということだ。

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

プライバシー保証は「平均」ではなく、一人ひとりへの約束である

医療 AI モデルが「プライバシー保護型」と呼ばれるとき、その主張は通常、一つの数字に基づいている。モデルの訓練に使われた患者全体で平均すると、ある個人のデータが訓練セットに含まれていたと推測される確率が低い、という数字だ。安心できそうに聞こえる。この論文の静かで居心地の悪い指摘は、それが見るべき数字ではないということだ。

プライバシーは平均ではない。各個人に対する約束だ。訓練データに入っていたことが、後になってその人を露出させないという約束である。そして平均値は、ほとんど全員にはその約束を守りながら、少数の人には完全に破ることができる。研究者らは、これまで使われてこなかった粒度で、患者一人ひとりのプライバシーリスクを測った。すると、まさにそれが見えた。集計すると安全に見えるモデルでも、特定個人については訓練データへの所属をほぼ完璧に漏らすことがある。そして漏らされる人は、不釣り合いなほど、もともと保護されにくい人たちだった。

著者らが行ったこと

Technical University of Munich の Moritz Knolle、Daniel Rückert、Hasso Plattner Institute の Georg Kaissis を中心に、Imperial College London の研究者らも加わったチームは、**メンバーシップ推論攻撃（membership inference attack）**というよく知られたプライバシー脅威について、問いを変えた。「データセット全体では平均して、この攻撃はどれくらい成功するか」ではなく、「この患者本人はどれくらい露出しているか」と尋ねた。

彼らはその患者単位の解析を、医用画像、心電図、電子健康記録という大きく異なるデータ型を含む**7 つの確立された医療データセット**で行い、それぞれについて多数のモデル（各データセットで 200 モデル）を調べた。各モデルの各患者について、攻撃者がその人の記録が訓練データに含まれていたかをどの程度確信を持って判別できるかを推定し、さらに疾患状態、自己申告人種、保険、性別、画像撮影プロトコルなどの集団別に結果を分解した。

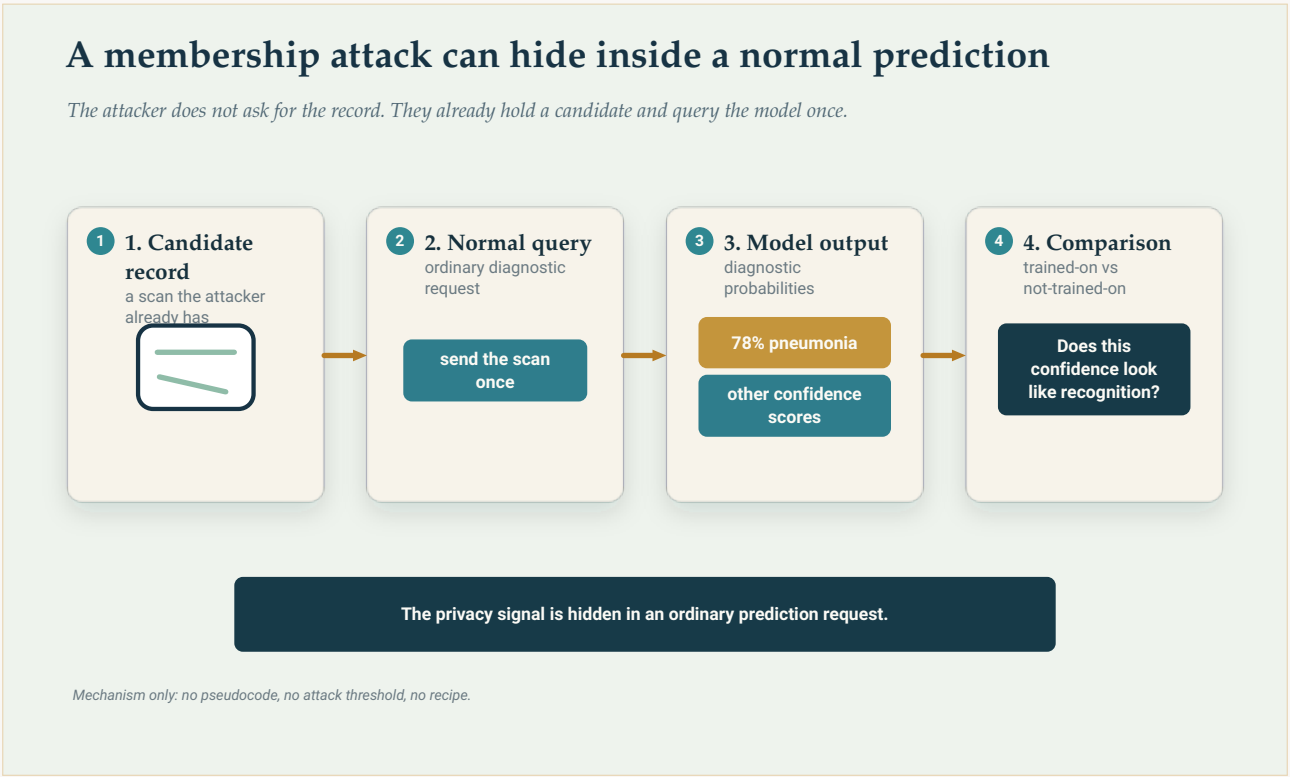
> **メンバーシップ推論攻撃とは実際に何をするのか**
>
> ここでの漏洩は、「モデルがあなたのカルテをそのまま吐き出す」というものよりずっと微妙だ。漏れるのは**訓練データへの含有情報** —— あなたのデータが訓練セットに含まれていたという事実そのもの —— である。
> まず、この種のモデルが普段何をするか考えよう。たとえば胸部 X 線画像を入力すると、肺炎 78%、心拡大、浮腫、浸潤影についてもそれぞれ確率を返す。この日常的な診断回答だけが、攻撃に使われる。特別な裏口は要らない。
> 利用される弱点はこうだ。モデルは普通、見たことのない例より、**訓練で実際に見た例に少しだけ自信を持ちすぎる**傾向がある。試験問題をこっそり事前に見た学生を想像するとよい。練習済みの問題では、答えが少し速く、少し自信満々に出てくる。その「癖」から、特定の記録がモデルの学習例の一つだったかを推測するのがメンバーシップ推論である。
> 攻撃は概念的にはこう進む。ある人の画像がモデルの訓練に使われたか知りたい攻撃者がい

る。モデルに「その記録を出せ」と頼むわけではない。候補となる画像 —— 本人の画像、あるいはそれに極めて近いもの —— はすでに持っている。それを通常の診断要求として一度入力し、モデルがどれほど確信を持った答えを返すかを見る。そして、その確信度が「その画像を訓練で見たモデル」と「見ていないモデル」のどちらに似ているかを比較する。攻撃者は普通のコンピュータで自前の代替モデル（reference model）を訓練し、訓練済み例に現れる特徴的な過剰確信を学ばせることで、この比較を比較的安価に行える。本物のモデルがその画像に異常なほど自信を示す —— まるで見覚えがあるかのように —— なら、訓練データに含まれていた強い手掛かりになる。

不気味なのは、要求そのものが攻撃に見えないことだ。臨床医が行うのと同じ診断問い合わせで、プライバシー信号は普通の予測の中に隠れている。だからこそリスクは具体的である。ただし現時点では、述べられた実験条件下で実証されたもので、現実世界で捕捉された攻撃ではない。

記録内容そのものが漏れないなら、所属だけがなぜ問題なのか。それは、**所属そのものが情報だから**だ。あるモデルが特定のがん免疫療法を受けた患者だけで訓練されていたとする。あなたの記録がそこに含まれていたと確認できれば、「あなたはおそらくそのがんだった」ということまで推測できる。保険会社や雇用主に推測されるべきではない種類の情報だ。中身は封じられたままでも、「そこに属していた」という事実が外へ出る。

著者らは、通常のモデル予測へのアクセス、候補記録、攻撃者自身の参照モデルという前提条件を明示している。攻撃手順を勧めるためではなく、脅威を曖昧にせず議論できるようにするためだ。



攻撃には特別なプライバシー API は要らない。モデルが過去に見た記録に異常に高い確信を示すなら、通常の診断問い合わせからでもメンバーシップ信号が漏れ得る。

Original Aurora diagram —The Clean Paper CC BY 4.0

分かったこと

結果は三つあり、後になるほど鋭い。

平均値は、露出している人を隠す。通常の方法どおり全体で集計すると、多くのモデルはかなり安心できるプライバシー性能に見える。大半の患者では攻撃は偶然以上に成功しない。しかし患者単位で見ると別の像が出た。Knolle は研究発表で、従来評価は「全患者にわたる平均リスクしか測ってこなかった。今回初めて患者個人レベルでリスクを調べると、まったく違う景色が見えた」と要約している。一部の個人では攻撃が**ほぼ完璧に**成功する。著者らの指標では攻撃 AUC が **0.95 以上**（0.5 はコイントス、1.0 は完全判別）だった。しかも、データセット全体の平均では偶然とほぼ変わらなく見える場合でもそうだった。

平均では偶然レベルに見えても、ごく一部の記録だけがはるかに識別されやすい「裾」として残り得る。この論文の要点は、プライバシーを集計値だけでなく患者単位でも確認しなければならないことだ。

Original Aurora diagram —The Clean Paper CC BY 4.0

露出は平等ではなく、少数派・希少例に集中する。ほぼ完全な攻撃に最も脆弱だった患者は、データ中で**過小代表されている集団**に体系的に属していた。少数民族、希少疾患の表現型、珍しい画像特徴などである。電子健康記録データセットでは、黒人患者は最脆弱記録の中に期待値より **31% 多く**現れた。マンモグラフィデータでは、悪性疑いとされた画像が危険な極端領域に **+1,179%** という大幅な過剰代表を示した。（この二つは例であって全体ではない。論文は複数の集団で同じ偏りを報告している。また、これらは小さな極端高リスク群の中での相対的な過剰代表を表す。黒人患者

や疑わしいマンモグラフィ患者の何割が露出している、という数字ではない。）モデルに似た例が少なければ、その患者記録は群れに紛れにくく、目立つ。そして「目立つこと」こそ、メンバーシップ攻撃が拾うものだ。プライバシー失敗はランダムではなく、すでに周縁にいる人へ集中する。

モデルが大きいほど悪化する。ほぼ完璧な攻撃にさらされる患者数は、モデル容量とともに大きく増えた。ある皮膚科データセットでは、最小モデルでほぼゼロだった割合が、最大モデルでは約 **10 人に 1 人**まで上昇した。医療 AI がより大きく、より高性能になっていく —— 分野全体が向かう方向だ —— につれ、この特定のリスクは軽くなるのではなく、深刻化すると著者らは警告する。

Rückert の要約は端的だ。「これは許容できるリスクではない。健康データは極めてセンシティブだ」。

「平均では安全」が間違った検査である理由

平均リスクが低ければ「問題なし」と読みたくなる。しかし、この論文の中心的な教訓は、ここで平均を取ることは単に粗いのではなく、**違うものを測っている**ということだ。

プライバシー保証に意味があるのは、中央値の人ではなく、最も危険にさらされる人にも成立するときだけだ。1000 人中 999 人は見分けられなくても、1 人だけほぼ完璧に識別できるモデルを、その 1 人にとって意味のある形で「99.9% プライベート」と呼ぶことはできない。そして、その 1 人が希少疾患患者や少数民族患者だと予測可能なら、指標は不完全なだけでなく、静かに差別的でもある。多数派の安全を測り、それを全員の安全と呼んでしまう。

この論文が迫る転換はそこだ。このモデルは平均でどれくらいプライベートか？ではなく、最も露出している患者は誰で、その人はどの集団に属するのか？。これは別の問いであり、プライバシーの問いとして本当に重要なのは後者だ。

この研究からは言えないこと

- 医療 AI を**やめるべきだ**とは言っていない。著者らの姿勢は撤退ではなく緩和だ。リスクを測り、直すのであって、開発停止を求めてはいない。
- すべての医療モデルが漏洩している、あるいは実際に運用中の特定モデルが攻撃された、という意味ではない。**リスクが存在し、不均等に分布している**こと、そして標準的な集計指標がそれを見逃すことを示している。
- あなたの記録が**すでに露出している**という意味ではない。攻撃には、モデルへのアクセス、候補記録、攻撃者側の参照モデルなど特定条件が必要で、気軽にできる能力ではない。
- 患者記録の**内容そのもの**が漏れることを示してはいない。漏れるのは、訓練データに含まれていたかどうかという情報であり、危険の種類が違う。
- 格差を一つの派手な数字へ還元してはいない。強く再現される結果は、集計指標が個人リスクを過小評価し、過小代表患者が最大の負担を受けるという**パターン**であり、それが複数データセット・数百モデルで見られたことだ。

証拠をどう評価するか

これは実証的・方法論的な結果で、かなり頑健だ。集計プライバシー指標が患者単位のリスクを体系的に過小評価すること、残るリスクが過小代表集団へ集中すること、モデル容量が大きいほど悪化すること —— このパターンは、**異なるデータ型を持つ 7 データセット**と、各データセットの多数のモデルで成立した。この広がりこそ、単一データセットだけのアーティファクトではなく、測定方法そのものに関する主張を可信にする。

ただし、二つの率直な注意点がある。第一に、これは著者らが構築した攻撃を実行して測るリスクの実証である。一定の仮定の下で能力のある攻撃者がどこまでできるかを特徴づけており、現実世界で攻撃がどれだけ頻繁に起きるかを測ってはいない。第二に、一部患者で攻撃がほぼ完璧に成功するという鮮烈な数字は、設計上もっとも不利な個人を表している。それが論文の目的そのものだ。読み方は「大多数が露出している」ではなく、「分布の極端な裾が平均から想像するよりはるかに重い」である。

適切な姿勢は、警鐘を鳴らしすぎることでも、無視することでもない。標準的な医療 AI プライバシー認証の方法が、自分自身の最悪ケースを見えなくしており、その最悪ケースが余裕の最も少ない患者へ偏ることを示した、慎重な複数データセット研究として読むべきだ。

なぜ重要なのか

これが単なる技術的な脚注ではない理由は二つある。

第一に、「プライバシー保護型」という言葉に何を許すべきかを変える。モデルが平均プライバシースコアを添えて公開され、その値が本当に低くても、メンバーシップ攻撃でほぼ完全に識別できる患者の一群を隠している可能性がある。論文の実務的な要求は具体的だ。公開前にプライバシーリスクを**患者単位**で評価すること、運用モデルに誰がアクセスできるか制御すること、そして**差分プライバシー（differential privacy）**のような技術を使うこと。差分プライバシーでは訓練中に小さく慎重に調整したノイズを加え、メンバーシップ攻撃を鈍らせる。ただしモデルの有用性との間には現実的なコストがあり、プライバシーと有用性のトレードオフは明示されるべきで、無料の対策ではない。「平均は確認した」だけでは、「確認した」と数えるべきではない。

第二に、プライバシーと公平性が結び付く。医療データで過小代表されているため、すでに医療 AI から十分な利益を受けにくい集団が、その AI によって最も弱くプライバシーを守られる集団でもある。一つ目の不平等を熱心に直そうとしている分野が、二つ目を別部署の問題として扱うことはできない。同じ人たちの話だからだ。

医療 AI のプライバシーについての安心できる物語 —— 測定した、平均は低い、だから問題ない —— を、この論文は分解する。技術から人を遠ざけるためではなく、基準を本来あるべき場所へ移すためだ。最も傷つきやすい人について確認して初めて、「約束を守っている」と言える。

まとめ

メンバーシップ推論攻撃は、特定の人の記録が AI モデルの訓練データに含まれていたかを推測する。所属そのものが、「その病気だった」といった情報を明らかにし得るため、元の記録内容が表に出なくても本物のプライバシー漏洩になる。従来研究は、こうした攻撃がデータセット全体で平均してどれくらい成功するかを測ってきた。この研究は、医用画像、ECG、電子健康記録を含む 7 つの医療データセットと多数のモデルで、患者一人ひとりのリスクを測った。その結果、集計指標はリスクを大きく過小評価することが分かった。データセット平均が安全に見えても、一部個人はほぼ完璧に識別できる（攻撃 AUC ≥ 0.95）。最も脆弱なのは、少数民族、希少疾患、珍しい画像特徴などの過小代表集団に体系的に属し、問題はモデルが大きくなるほど悪化する。著者らは医療 AI を捨てるよう求めてはいない。患者単位でプライバシーを測り、モデルへのアクセスを制御し、差分プライバシーを使うよう求めている。結論は、「平均的には安全」というだけではプライバシー保証にならず、その失敗はすでに保護の薄い患者に集中する、ということだ。

冷静に見ると

論文が示したこと： 7 つの医療データセットと多数のモデルを通じて、患者単位のメンバーシップ推論リスクは集計指標が示すより一部個人で大幅に高く、攻撃 AUC ≥ 0.95 というほぼ完全な判別に達することがある。その残余リスクは過小代表集団に不釣り合いに集中し、モデル容量とともに増える。

もっともらしいが、まだ証明されていないこと： 現実の攻撃者がすでに運用中の臨床モデルに対してこの方法を使っていること。この研究が示したのは、明示された仮定の下での攻撃能力とその分布であり、野外での攻撃頻度ではない。

示していないこと： 医療 AI をやめるべきであること。すべてのモデルが漏洩していること、または特定の運用モデルが侵害されたこと。患者記録の内容がメンバーシップではなく直接露出すること。一つの数値だけで表せる格差 —— 頑健なのは単一数字ではなくパターンである。

主な限界： 実際の侵害事例ではなく、著者ら自身の攻撃で最悪ケースのリスクを測っている。派手な数字は設計上、最も露出した個人を表す。集団ごとの正確な格差は単一値へまとめられず、複数データセットに一貫して現れるパターンとして示される。

一般読者はどの程度信頼すべきか？ 集計プライバシー指標が個人リスクを過小評価すること、残余リスクが過小代表患者へ集中すること、それがモデルサイズとともに悪化することには高い確信を持ってよい。多くのデータセットとモデルで示されている。現実世界でこの種の攻撃がどれほど頻繁かについては中程度以下の確信にとどめるべきで、この研究はそこを測っていない。安全な読み方は「医療 AI があなたのデータを漏らしている」でも「プライバシー問題は解決済み」でもない。「標準的なプライバシー確認は自分の最悪ケースを見逃し、そのケースは最も脆弱な患者に偏る。だから確認方法を変える必要がある」である。

出典

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