The Clean Paper

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

Une IA medicale peut etre privee en moyenne et exposer quand meme certains patients - surtout les sous-representes

Un modele d'IA medicale dit protecteur de la vie privee repose souvent sur une moyenne. Cette etude soutient que c'est le mauvais test. En etudiant des attaques d'inference d'appartenance - qui revelent si le dossier d'une personne precise etait dans les donnees d'entrainement d'un modele, et peuvent donc trahir une maladie - les auteurs mesurent le risque patient par patient, pas en agregat, sur sept jeux de donnees medicaux et de nombreux modeles. Des modeles rassurants en moyenne peuvent laisser identifier certains individus presque parfaitement (AUC d'attaque >= 0,95); les exposes viennent systematiquement de groupes sous-representes; et le risque augmente avec la taille des modeles. Le message n'est pas d'abandonner l'IA medicale, mais de mesurer la confidentialite au niveau individuel, controler l'acces aux modeles et utiliser la confidentialite differentielle.

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

Une garantie de confidentialité est une promesse faite à une personne, pas à une moyenne

Quand un modèle d'IA médicale est présenté comme « protecteur de la vie privée », l'affirmation repose souvent sur un nombre agrégé : parmi tous les patients dont les données ont servi à l'entraînement, le risque moyen qu'un attaquant déduise l'appartenance d'un individu au jeu de données est faible. Cela paraît rassurant. Le point discret — et inconfortable — de cet article est que ce n'est pas nécessairement le bon nombre à regarder.

La confidentialité n'est pas une moyenne. C'est une promesse faite à chaque individu : le fait que ses données aient participé à l'entraînement ne devrait pas permettre de l'exposer. Or une moyenne peut être rassurante pour presque tout le monde tout en masquant un risque très élevé pour quelques personnes. Les chercheurs ont donc mesuré ce risque patient par patient, avec une résolution inhabituelle, et trouvent précisément ce scénario : des modèles qui paraissent sûrs en moyenne peuvent révéler presque parfaitement l'appartenance de certains individus — et ces individus appartiennent de façon disproportionnée à des groupes déjà sous-représentés.

Ce que les auteurs ont fait

L'équipe menée par Moritz Knolle, avec Daniel Rückert, Georg Kaissis et leurs collègues, est partie d'une menace connue : l'**attaque par inférence d'appartenance**. Mais au lieu de demander « en moyenne, à quelle fréquence l'attaque réussit-elle ? », elle a posé une question individuelle : « pour *ce patient précis*, quel est le niveau de risque ? »

Les auteurs ont mené cette analyse sur **sept jeux de données médicaux établis**, couvrant l'imagerie, les électrocardiogrammes et les dossiers de santé électroniques, avec 200 modèles entraînés par jeu de données. Pour chaque patient et chaque modèle, ils ont estimé avec quelle confiance un attaquant pouvait décider si le dossier avait participé ou non à l'entraînement. Ils ont ensuite décomposé ces résultats selon différents groupes : maladie, race autodéclarée, assurance, sexe et protocole d'imagerie.

> **Ce qu'est réellement une attaque par inférence d'appartenance**
>
> La fuite étudiée ici est plus subtile que « le modèle recraché votre dossier ». Elle porte sur *l'appartenance* : le simple fait que vos données aient fait partie du jeu d'entraînement.
> Un modèle médical ordinaire reçoit, par exemple, une radiographie thoracique et renvoie des probabilités : pneumonie, œdème, consolidation. L'attaque n'a besoin de rien de plus exotique que cette sortie diagnostique normale.
> La faiblesse exploitée est la suivante : un modèle peut être légèrement **plus confiant** sur les exemples exacts qu'il a vus pendant l'entraînement que sur des exemples nouveaux. C'est un peu comme un étudiant qui aurait déjà vu certaines questions d'examen : ses réponses sont parfois anormalement assurées. Déterminer, à partir de ce signal, si un dossier donné faisait partie de l'entraînement constitue une inférence d'appartenance.

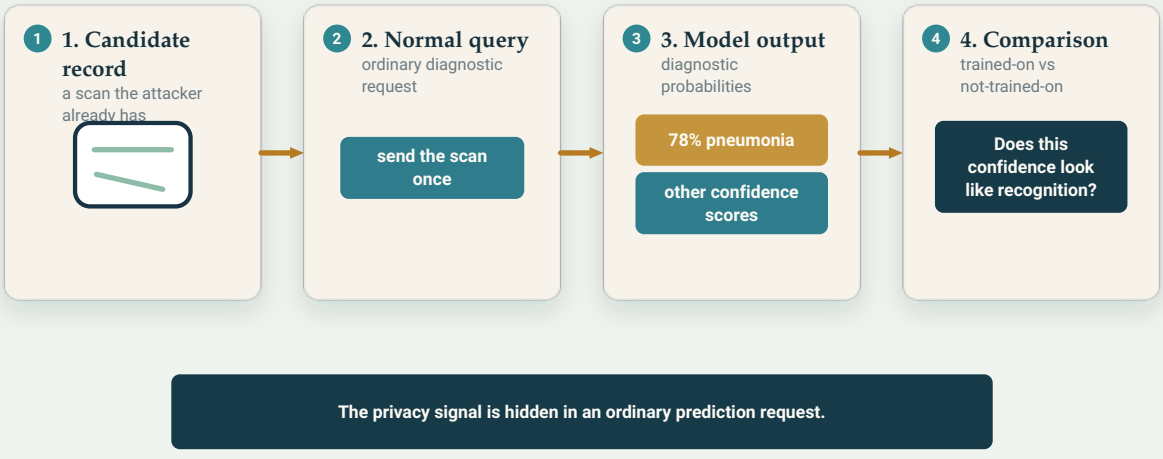
Concrètement, un attaquant qui souhaite savoir si le dossier d'une personne a servi à entraîner le modèle possède déjà un exemple candidat — par exemple une image médicale. Il l'envoie comme une requête diagnostique ordinaire, observe la confiance du modèle, puis compare ce comportement à celui attendu d'un modèle ayant vu ou non cet exemple. Cette comparaison peut être apprise à l'aide de modèles de référence. Une confiance anormalement élevée peut alors constituer un signal d'appartenance.

Le point préoccupant est que la requête ne ressemble pas nécessairement à une attaque : elle peut être identique à une requête clinique normale. Le signal de confidentialité est caché dans une prédiction ordinaire. Le risque est donc concret, même si l'article le démontre dans des conditions expérimentales et non à partir d'attaques observées dans la nature.

Pourquoi l'appartenance compte-t-elle si le contenu du dossier n'est pas révélé ? Parce que *l'appartenance est elle-même une information*. Si un modèle a été entraîné exclusivement sur des patients recevant une immunothérapie contre un cancer, établir que votre dossier appartenait à cet ensemble peut déjà révéler une information médicale sensible. Le contenu du dossier reste fermé ; le fait d'en faire partie peut tout de même s'échapper.

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.



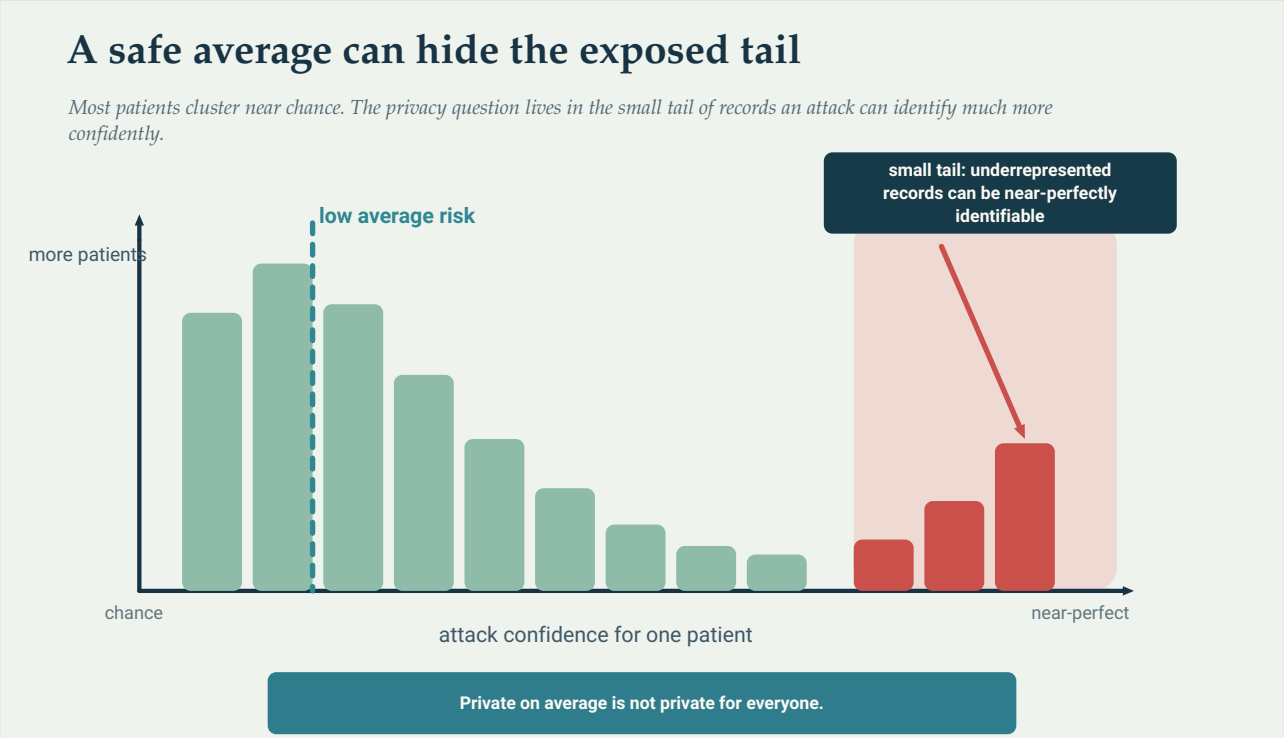
Mechanism only: no pseudocode, no attack threshold, no recipe.

L'attaque n'a pas besoin d'une interface spéciale consacrée à la confidentialité. Une requête diagnostique ordinaire peut révéler un signal d'appartenance si le modèle se montre anormalement confiant sur un dossier déjà vu.

Ce qu'ils ont trouvé

Trois résultats se dégagent.

Les moyennes peuvent masquer les personnes les plus exposées. Mesurés de manière agrégée, de nombreux modèles paraissent rassurants : pour la plupart des patients, l'attaque ne fait guère mieux que le hasard. Mais l'analyse individuelle raconte une autre histoire. Pour certains dossiers, l'attaque réussit **presque parfaitement** — les auteurs utilisent notamment une AUC d'attaque de **0,95 ou plus**, où 0,5 correspond au hasard et 1,0 à une séparation parfaite — alors même que la moyenne du jeu de données reste proche du hasard.



La moyenne peut rester proche du hasard alors qu'une petite fraction de dossiers est beaucoup plus exposée. L'idée centrale de l'article est que la confidentialité doit être évaluée au niveau du patient, pas seulement en moyenne.

L'exposition est inégalement répartie et touche davantage les groupes sous-représentés. Les patients les plus vulnérables aux attaques presque parfaites appartiennent systématiquement davantage à des groupes **sous-représentés dans les données** : minorités ethniques, phénotypes rares ou caractéristiques d'imagerie inhabituelles. Dans le jeu de dossiers électroniques, les patients noirs apparaissent **31 % plus souvent que prévu** parmi les dossiers les plus vulnérables ; dans le jeu de mammographie, les images suspectes de malignité sont surreprésentées dans cette zone extrême de **+1 179 %**. Ces chiffres décrivent une surreprésentation relative dans une petite queue de distribution, pas la fraction de tous les patients de ces groupes qui seraient exposés. Une interprétation possible est qu'un modèle disposant de moins d'exemples semblables « mémorise » plus facilement certains

dossiers atypiques.

Les modèles plus grands aggravent le problème. Le nombre de patients exposés à des attaques presque parfaites augmente avec la capacité du modèle. Dans un jeu dermatologique, la proportion passe de presque zéro pour le plus petit modèle à environ **un patient sur dix** pour le plus grand. À mesure que l'IA médicale se tourne vers des modèles plus volumineux, ce risque précis peut donc augmenter plutôt que diminuer.

Le résumé de Daniel Rückert est sans détour : « This is not a tolerable risk. Health data is highly sensitive. »

Pourquoi « sûr en moyenne » est le mauvais test

Il est tentant de lire un faible risque moyen comme un certificat de sécurité. La leçon centrale de l'article est que la moyenne n'est pas seulement imprécise : elle répond à une question différente.

Une garantie de confidentialité n'a de sens que si elle protège aussi les personnes les plus exposées. Un modèle pour lequel 999 patients sur 1 000 sont difficiles à identifier mais où un seul patient peut être reconnu presque parfaitement n'est pas « privé à 99,9 % » pour cette personne. Et si les personnes les plus exposées appartiennent de façon prévisible à des groupes minoritaires ou à des patients atteints de maladies rares, la moyenne masque aussi une dimension d'équité.

Le déplacement proposé par l'article est donc simple : passer de « *à quel point ce modèle est-il privé en moyenne ?* » à « *qui est le plus exposé, et à quel point ?* ». Les deux questions ne sont pas équivalentes, et la seconde est celle qui correspond réellement à une promesse faite à une personne.

Ce que cela ne prouve pas

- L'article ne dit **pas** qu'il faut abandonner l'IA médicale. Son objectif est de mieux mesurer et réduire le risque.
- Il ne signifie **pas** que chaque modèle médical fuit des informations, ni qu'un système clinique précis a déjà été attaqué de cette manière. Il montre qu'un risque peut exister et être réparti de façon très inégale.
- Il ne signifie **pas** que vos dossiers sont déjà exposés. L'attaque suppose des conditions précises : accès au modèle, possession d'un dossier candidat et capacité à construire une référence.
- Il ne montre **pas** que le contenu détaillé du dossier est révélé. La fuite porte sur l'appartenance au jeu d'entraînement, qui peut néanmoins être sensible.
- Il ne réduit **pas** les disparités de confidentialité à un seul chiffre universel. Le résultat robuste est le motif observé : les moyennes sous-estiment le risque individuel et les groupes sous-représentés supportent davantage la queue de risque.

Quelle est la solidité de la preuve ?

Le résultat est méthodologiquement et empiriquement solide. Le même motif — le risque individuel est beaucoup plus hétérogène que ne le suggère la moyenne, les groupes sous-représentés concentrent davantage le risque résiduel et les modèles plus grands aggravent l'exposition — apparaît sur **sept jeux de données de natures différentes** et avec de nombreux modèles. Cette diversité rend le résultat plus convaincant qu'une anomalie propre à un seul ensemble de données.

Deux réserves restent importantes. D'abord, il s'agit d'une démonstration de *risque* obtenue avec les attaques construites par les auteurs. Elle caractérise ce qu'un attaquant capable *pourrait* faire dans certaines conditions, pas la fréquence réelle de telles attaques dans les systèmes déployés. Ensuite, les nombres les plus spectaculaires concernent par définition les individus situés dans l'extrême de la distribution. C'est précisément le sujet de l'étude, mais il faut les lire comme « la queue est beaucoup plus dangereuse que la moyenne ne le laisse penser », et non comme « la plupart des patients sont exposés ».

La lecture appropriée n'est donc ni la panique ni le rejet de la technologie. C'est une étude large montrant que la façon habituelle de résumer la confidentialité d'un modèle médical peut ignorer ses cas les plus vulnérables.

Pourquoi c'est important

Deux éléments rendent ce résultat plus qu'un détail technique.

D'abord, l'article change ce que l'expression « respectueux de la vie privée » devrait être autorisée à signifier. Une moyenne rassurante peut cacher un sous-ensemble de patients presque parfaitement identifiables. La conséquence pratique est concrète : évaluer le risque **patient par patient** avant publication, contrôler l'accès aux modèles déployés et, lorsque c'est approprié, utiliser des techniques comme la **confidentialité différentielle**, qui ajoute un bruit mathématiquement calibré pendant l'entraînement afin de limiter ce type d'attaque au prix d'un compromis sur l'utilité. « La moyenne est faible » ne devrait plus suffire comme justification.

Ensuite, l'article relie confidentialité et équité. Les groupes sous-représentés dans les données médicales — souvent déjà ceux pour lesquels les modèles sont les moins performants — peuvent aussi être ceux dont la confidentialité est la plus fragile. Il ne s'agit donc pas nécessairement de deux problèmes indépendants : ils peuvent toucher les mêmes personnes.

Le récit rassurant — *nous avons calculé la moyenne, elle est faible, donc tout va bien* — est précisément celui que l'article remet en cause. Non pour condamner l'IA médicale, mais pour déplacer le standard là où il devrait être : une promesse de confidentialité ne peut être considérée comme tenue qu'après avoir vérifié les personnes les plus susceptibles d'être lésées.

Résumé clair

Les attaques par inférence d'appartenance cherchent à déterminer si le dossier d'une personne précise faisait partie des données d'entraînement d'un modèle. Comme cette appartenance peut à elle seule révéler une information médicale, elle constitue une véritable fuite de confidentialité même lorsque le dossier n'est jamais reproduit. Les travaux antérieurs mesuraient surtout le succès moyen de ces attaques. Cette étude l'évalue *patient par patient* sur sept jeux de données médicaux et de nombreux modèles. Elle montre que les moyennes peuvent fortement sous-estimer la queue du risque : certains individus sont presque parfaitement identifiables (AUC $\geq$ 0,95) alors que la moyenne paraît proche du hasard ; les personnes les plus vulnérables proviennent davantage de groupes sous-représentés ; et le risque augmente avec la taille du modèle. Les auteurs ne proposent pas d'abandonner l'IA médicale, mais de mesurer le risque au niveau individuel, de contrôler l'accès et d'utiliser des protections comme la confidentialité différentielle. Point central : « privé en moyenne » n'est pas une garantie individuelle.

Mise au point

Ce que l'article montre : sur sept jeux de données médicaux et de nombreux modèles, le risque individuel d'inférence d'appartenance varie fortement entre patients. Certains dossiers peuvent atteindre une AUC d'attaque $\geq$ 0,95 alors que la moyenne du jeu de données paraît rassurante ; ce risque résiduel touche davantage des groupes sous-représentés et augmente avec la capacité du modèle.

Ce qui est plausible mais non démontré : que des attaquants exploitent déjà ce mécanisme contre des modèles cliniques déployés. L'étude démontre une capacité d'attaque et sa distribution, pas sa fréquence dans le monde réel.

Ce que cela ne montre pas : qu'il faut abandonner l'IA médicale ; que chaque modèle fuit ; qu'un système précis a été compromis ; que le contenu complet des dossiers est révélé ; ou qu'il existe un chiffre unique résumant toutes les disparités.

Principales limites : le risque est mesuré au moyen d'attaques expérimentales, pas d'incidents observés ; les chiffres les plus élevés décrivent intentionnellement les pires cas ; les disparités sont établies comme un motif récurrent plutôt que comme une valeur universelle.

Niveau de confiance pour un lecteur général : élevé sur le fait que les moyennes peuvent sous-estimer fortement le risque individuel, que ce risque résiduel se concentre davantage sur les patients sous-représentés et qu'il peut augmenter avec la taille du modèle. Plus modéré sur la fréquence réelle des attaques. La bonne lecture n'est ni « l'IA médicale divulgue vos données » ni « la confidentialité est résolue », mais « le test moyen peut être aveugle à ses pires cas ».

Sources

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