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 · 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 confidentialite est une promesse faite a une personne, pas a une moyenne

Quand un modele d'IA medicale est appele "protecteur de la vie privee", l'affirmation repose souvent sur un nombre : sur tous les patients dont les donnees ont servi a l'entrainement, la probabilite moyenne d'inferer l'appartenance d'un individu est faible. Cela semble rassurant. Le point discret et inconfortable de cet article est que ce n'est pas le bon nombre.

La confidentialite n'est pas une moyenne. C'est une promesse faite a chaque individu : le fait d'etre dans le jeu d'entrainement ne doit pas revenir l'exposer. Or une moyenne peut tenir cette promesse pour presque tout le monde tout en la brisant completement pour quelques personnes. Les chercheurs ont mesure le risque patient par patient, avec une resolution encore inedite, et trouvent exactement cela : des modeles qui paraissent surs en agregat peuvent reveler l'appartenance de certains individus presque parfaitement - et ces individus sont, disproportionnellement, deja les moins proteges.

Ce que les auteurs ont fait

L'equipe menee par Moritz Knolle, avec Daniel Rückert, Georg Kaissis et leurs collegues, a pris une menace connue, l'**attaque d'inference d'appartenance**, et a change la question. Au lieu de demander "en moyenne, a quelle frequence l'attaque reussit-elle dans le dataset ?", ils ont demande "pour *ce patient precis*, a quel point est-il expose ?"

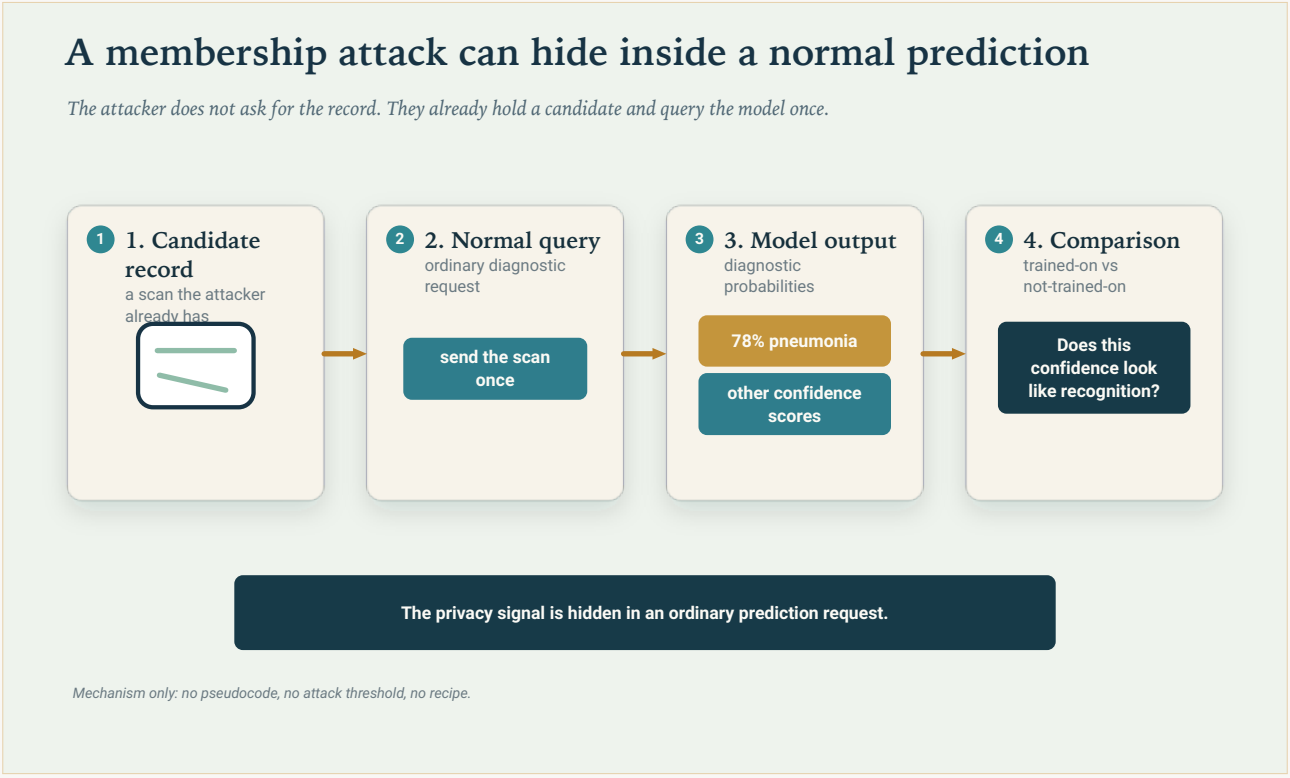
Ils ont mene cette analyse par patient sur **sept jeux de donnees medicaux etablis**, couvrant imagerie, electrocardiogrammes et dossiers de sante electroniques, avec 200 modeles par jeu. Pour chaque patient dans chaque modele, ils ont estime avec quelle confiance un attaquant pouvait dire si son dossier avait fait partie de l'entrainement, puis ont decompose les resultats par groupe : maladie, race auto-declaree, assurance, sexe et protocole d'imagerie.

> **Ce qu'est vraiment une attaque d'inference d'appartenance**
>
> La fuite ici est plus subtile que "le modele recrache votre dossier". Elle porte sur *l'appartenance* : le simple fait que vos donnees etaient dans l'ensemble d'entrainement.
> Un modele medical ordinaire recoit une radio thoracique et renvoie des probabilites : pneumonie, oedeme, consolidation. Cette reponse diagnostique normale est tout ce que l'attaque utilise. Rien d'exotique.
> La faiblesse est la suivante : un modele est souvent un peu **plus confiant** sur les exemples exacts qu'il a vus a l'entrainement que sur ceux qu'il n'a jamais vus. Comme un etudiant qui aurait vu l'examen avant : les reponses viennent trop vite, trop surement. Determiner, a partir de ce signe, si un dossier particulier faisait partie des exemples etudies, c'est l'inference d'appartenance.
> L'attaque est donc simple. Quelqu'un veut savoir si le scan d'une personne a servi a entrainer

le modele. Il ne demande pas le dossier au modele; il possede deja un scan candidat. Il l'envoie comme une requete diagnostique ordinaire et observe la confiance de la reponse. Puis il compare cette confiance a ce qu'il attend d'un modele qui aurait vu ce scan ou non, comparaison qu'il peut apprendre avec un modele de reference. Si le vrai modele est anormalement sur de ce scan, comme s'il le reconnaissait, c'est une preuve forte d'appartenance.

Le point inquietant est que la requete ne ressemble pas a une attaque : c'est la meme requete diagnostique qu'un clinicien pourrait faire, et le signal de confidentialite est cache dans une prediction ordinaire. Le risque est concret meme si, jusqu'ici, il a ete demontre sous hypotheses de laboratoire et non observe en attaque sauvage.

Pourquoi l'appartenance compte-t-elle si le contenu du dossier ne fuit pas ? Parce que *l'appartenance est un fait*. Si un modele a ete entraine sur des patients ayant recu une immunotherapie contre un cancer, confirmer que votre dossier en fait partie revele probablement ce cancer. Le contenu reste scelle; le fait d'appartenir s'echappe.

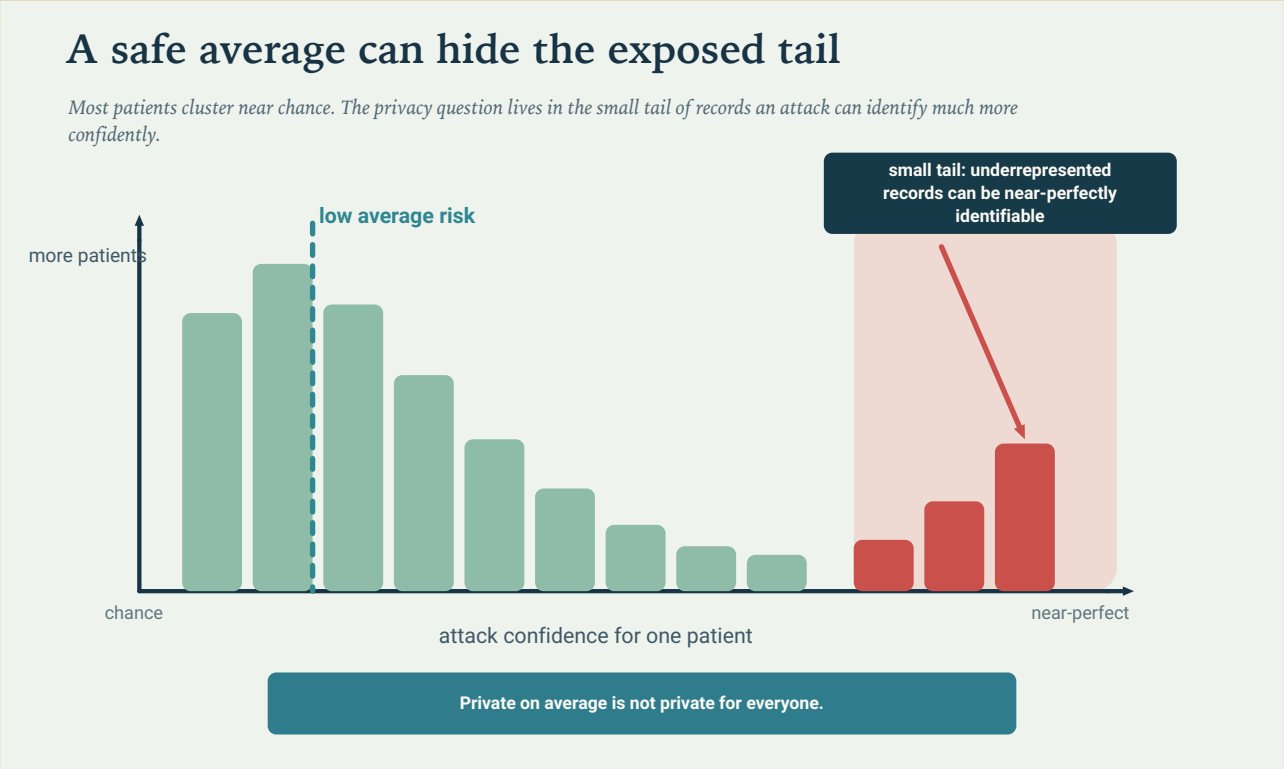


L'attaque n'a pas besoin d'une API speciale de confidentialite. Une requete diagnostique normale peut laisser fuir un signal d'appartenance si le modele est anormalement confiant sur un dossier deja vu.

Ce qu'ils ont trouve

Trois resultats, de plus en plus nets.

Les moyennes cachent les personnes exposees. Mesures en agregat, comme d'habitude, beaucoup de modeles semblent rassurants : l'attaque ne fait pas mieux que le hasard pour la plupart des patients. La vue patient par patient raconte autre chose. Pour certains individus, l'attaque reussit **presque parfaitement** - les auteurs le mesurent comme une AUC d'attaque de **0,95 ou plus** (0,5 est le hasard, 1,0 parfait) - meme quand la moyenne du dataset ressemble au hasard.



La moyenne peut rester pres du hasard tandis qu'une petite queue de dossiers demeure beaucoup plus exposee. Le point de l'article est que la confidentialite doit etre verifiee au niveau du patient, pas seulement en agregat.

L'exposition est inegale - et touche les sous-representes. Les patients les plus vulnerables aux attaques presque parfaites appartiennent systematiquement a des groupes **sous-representes dans les donnees** : minorites ethniques, phenotypes rares, caracteristiques d'imagerie inhabituelles. Dans le jeu de dossiers electroniques, les patients noirs apparaissent **31 % plus souvent** que prevu parmi les dossiers les plus vulnerables; dans la mammographie, les scans suspects de malignite sont surrepresentes dans cette zone de danger de **+1 179 %**. Ces nombres sont des exemples et des surrepresentations relatives dans la petite queue extreme, pas la fraction de tous ces patients qui sont exposes. Un modele a moins d'exemples similaires pour diluer ces patients; ils ressortent, et ressortir est exactement ce qu'une attaque detecte.

Les grands modeles aggravent le probleme. Le nombre de patients exposes a des attaques presque parfaites augmente fortement avec la capacite du modele. Dans un jeu dermatologique, la part passe de presque zero dans le plus petit modele a environ **un sur dix** dans le plus grand. Alors que l'IA medicale va vers des modeles plus grands, ce risque specifique augmente, pas l'inverse.

Le resume de Rückert est direct : “This is not a tolerable risk. Health data is highly sensitive.”

Pourquoi “sûr en moyenne” est le mauvais test

La tentation est de lire un faible risque moyen comme un certificat de securite. La lecon centrale est que la moyenne n’est pas seulement imprecise : elle mesure la mauvaise chose.

Une garantie de confidentialite n’a de sens que si elle tient pour la personne la plus exposee, pas pour la personne au milieu. Un modele ou 999 patients sur 1 000 sont inidentifiables mais ou un seul peut etre identifie presque parfaitement n’est pas “99,9 % prive” pour cette personne - et si cette personne est de facon previsible le patient a maladie rare ou le patient d’une minorite, la metrique devient discriminatoire. Elle signale la securite de la majorite et l’appelle securite de tous.

Le changement impose par l’article est donc celui-ci : passer de *a quel point ce modele est-il prive en moyenne ?* a *qui est le patient le plus expose, et qui est-il ?* Ce sont deux questions differentes, et seule la seconde est une question de confidentialite.

Ce que cela ne prouve pas - ni ne pretend

- Cela ne dit **pas** qu’il faut abandonner l’IA medicale. Le cadrage est la mitigation, pas le retrait.
- Cela ne veut **pas** dire que chaque modele medical fuit, ni qu’un modele deployee donne a deja ete attaque. Cela montre que le *risque existe et est inegalement distribue*, et que les metriques agregees le manquent.
- Cela ne veut **pas** dire que vos dossiers sont deja exposes. L’attaque demande des conditions precises : acces au modele, dossier candidat et infrastructure de reference.
- Cela ne montre **pas** que le *contenu* des dossiers fuit. Ce qui fuit est l’appartenance, dangereuse pour une autre raison.
- Cela ne reduit **pas** les disparites a un seul nombre. Le resultat robuste est le *motif* : les moyennes sous-estiment le risque individuel, et les patients sous-representes le portent davantage.

Quelle est la force de la preuve ?

C’est un resultat empirique et methodologique robuste. Le motif - les metriques agregees sous-estiment le risque patient par patient, le risque residuel se concentre sur les groupes sous-representes et empire avec la capacite - tient sur **sept jeux de donnees de types differents** et beaucoup de modeles. Cette largeur rend la mesure credible plutot qu’anecdotique.

Deux reserves. D’abord, c’est une demonstration de *risque*, mesuree par les attaques construites par les auteurs; elle caracterise ce qu’un attaquant capable *pourrait* faire sous hypotheses, pas la frequence des attaques reelles. Ensuite, les nombres les plus frappants decrivent les individus les plus

exposes *par construction*; c'est le point meme de l'etude, et il faut les lire comme "la queue est bien plus lourde que la moyenne ne le dit", pas "la plupart des patients sont exposes".

La bonne posture n'est ni panique ni rejet : une etude multi-dataset prudente montrant que la facon standard de certifier la confidentialite de l'IA medicale est aveugle a ses pires cas, et que ces pires cas tombent sur les patients qui ont le moins de marge.

Pourquoi c'est important

Deux choses rendent cela plus qu'un detail technique.

D'abord, l'article change ce que "privacy-preserving" devrait avoir le droit de vouloir dire. Un score moyen peut etre bas et cacher quand meme un sous-ensemble de patients presque parfaitement identifiables. La demande pratique est concrete : evaluer le risque **par patient** avant publication, controler l'acces aux modeles deployes et utiliser des techniques comme la **confidentialite differentielle** - du bruit soigneusement calibre ajoute a l'entrainement pour affaiblir les attaques, avec un vrai compromis confidentialite-utilite. "Nous avons verifie la moyenne" ne devrait plus suffire.

Ensuite, cela tresse confidentialite et equite. Les memes groupes sous-representes dans les donnees medicales, donc deja moins bien servis par l'IA medicale, sont ceux dont la confidentialite est le moins protegee. Ce ne sont pas deux problemes separes; ce sont les memes personnes.

Le recit rassurant - *nous avons mesure, la moyenne est basse, tout va bien* - est celui que l'article defait. Non pour faire fuir la technologie, mais pour deplacer le standard la ou il doit etre : une promesse que l'on ne peut pretendre tenir qu'apres l'avoir verifiee pour la personne la plus susceptible d'etre blessee.

Resume clair

Les attaques d'inference d'appartenance cherchent a savoir si le dossier d'une personne precise faisait partie des donnees d'entrainement d'un modele. Comme cette appartenance peut reveler une maladie, c'est une vraie fuite de confidentialite meme si le dossier lui-meme ne sort jamais. Les travaux precedents mesuraient la reussite moyenne de ces attaques. Cette etude la mesure *par patient*, sur sept jeux de donnees medicaux et de nombreux modeles, et montre que les metriques agregees sous-estiment fortement le risque : certains individus sont identifiables presque parfaitement (AUC >= 0,95) meme quand la moyenne semble sure; les plus vulnerables viennent de groupes sous-representes; et le probleme augmente avec la taille du modele. Les auteurs ne disent pas d'abandonner l'IA medicale, mais de mesurer au niveau individuel, controler l'acces et utiliser la confidentialite differentielle. Point central : "prive en moyenne" n'est pas une garantie de confidentialite.

No-BS check

Ce que l'article montre : Sur sept jeux de donnees medicaux et de nombreux modeles, le risque d'inference d'appartenance par patient est beaucoup plus eleve pour certains individus que ne le suggerent les metriques agregees, jusqu'a une AUC >= 0,95, et ce risque residuel touche disproportionnellement les groupes sous-representes et augmente avec la capacite du modele.

Ce qui est plausible mais non prouve : Que des attaquants exploitent deja cela contre des modeles cliniques deployes; l'etude demontre la capacite et sa distribution sous hypotheses, pas la frequence reelle.

Ce que cela ne montre pas : Qu'il faut abandonner l'IA medicale; que chaque modele fuit; qu'un modele specifique a ete compromis; que le contenu des dossiers fuit; un seul chiffre universel de disparite.

Limites principales : Risque mesure via les attaques des auteurs, pas incidents observes; chiffres dramatiques decrivant les pires cas par design; disparites montrees comme motif coherent plutot que reduites a un nombre.

Confiance pour un lecteur general : Haute que les moyennes sous-estiment le risque individuel, que le risque residuel se concentre sur les patients sous-representes et que cela empire avec la taille. Moyenne sur la frequence reelle des attaques. Lecture sure : ni "l'IA medicale fuit vos donnees", ni "la confidentialite est resolue", mais "le test standard est aveugle a 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-ai

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