

Eine medizinische KI kann im Durchschnitt privat sein und dennoch einzelne Patienten entblößen — besonders die Unterrepräsentierten

Bei einem medizinischen KI-Modell, das als „datenschutzwahrend" gilt, beruht diese Aussage meist auf einer durchschnittlichen Kennzahl. Diese Studie argumentiert, dass das der falsche Test ist. Die Autoren untersuchten Membership-Inference-Angriffe — Angriffe, die verraten können, ob der Datensatz einer bestimmten Person im Training enthalten war und damit beispielsweise indirekt eine Erkrankung offenlegen — und maßen das Risiko nicht aggregiert, sondern für jeden einzelnen Patienten. Über sieben medizinische Datensätze mit Bildgebung, EKG und elektronischen Gesundheitsakten sowie viele Modelle hinweg zeigte sich: Modelle, die im Durchschnitt sicher aussehen, können einzelne Personen nahezu perfekt identifizierbar machen (Angriffs-AUC $\geq 0{,}95$); betroffen sind systematisch Angehörige unterrepräsentierter Gruppen wie ethnische Minderheiten, Patienten mit seltenen Krankheiten oder ungewöhnlichen Bildmerkmalen; und mit wachsender Modellgröße wird das Problem stärker. Die Autoren fordern nicht den Verzicht auf medizinische KI, sondern Datenschutzmessungen pro Patient, kontrollierten Modellzugriff und Differential Privacy. Der unbequeme Kern: „im Durchschnitt privat" ist keine Datenschutzgarantie — und sie versagt gerade bei Patienten, die ohnehin am wenigsten geschützt sind.

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

Eine Datenschutzgarantie ist ein Versprechen an einen Menschen, nicht an einen Durchschnitt

Wenn ein medizinisches KI-Modell als „datenschutzwahrend“ bezeichnet wird, stützt sich diese Aussage meist auf eine Zahl: Über alle Patienten hinweg, deren Daten das Modell trainiert haben, ist die durchschnittliche Wahrscheinlichkeit gering, dass sich die Zugehörigkeit einer einzelnen Person zum Trainingsdatensatz erschließen lässt. Das klingt beruhigend. Der leise, unbequeme Punkt dieser Arbeit ist: Es ist die falsche Zahl.

Datenschutz ist kein Durchschnitt. Er ist ein Versprechen an jeden Einzelnen — dass die Tatsache, im Trainingsdatensatz gewesen zu sein, später nicht gegen ihn verwendet werden kann. Ein Durchschnitt kann dieses Versprechen für fast alle einhalten und es für wenige vollständig brechen. Die Forschenden maßen das Datenschutzrisiko Patient für Patient, in einer Auflösung, die so zuvor nicht verwendet worden war, und fanden genau das: Modelle, die insgesamt sicher aussehen, können bei bestimmten Personen die Mitgliedschaft im Trainingsdatensatz nahezu perfekt verraten — und überproportional betroffen sind gerade diejenigen, die ohnehin am wenigsten geschützt sind.

Was die Autoren gemacht haben

Das Team — unter Leitung von Moritz Knolle, mit Daniel Rückert (beide Technische Universität München) und Georg Kaissis (Hasso-Plattner-Institut) sowie Kolleginnen und Kollegen vom Imperial College London — nahm eine bekannte Datenschutzbedrohung, den **Membership-Inference-Angriff**, und änderte die Fragestellung. Statt zu fragen: „Wie oft gelingt dieser Angriff im Durchschnitt über den gesamten Datensatz?“, fragten sie: „Wie stark ist *dieser konkrete Patient* gefährdet?“

Sie führten diese patientenbezogene Analyse auf **sieben etablierten medizinischen Datensätzen** durch, die sehr unterschiedliche Datentypen umfassen — medizinische Bildgebung, Elektrokardiogramme und elektronische Gesundheitsakten — und untersuchten dabei jeweils 200 Modelle pro Datensatz. Für jeden Patienten in jedem Modell schätzten sie, wie sicher ein Angreifer erkennen könnte, ob der Datensatz dieser Person zum Training gehört hatte. Anschließend zerlegten sie die Ergebnisse nach Gruppen: Krankheitsstatus, selbst angegebene ethnische Zugehörigkeit, Versicherung, Geschlecht und Bildgebungsprotokoll.

Was ein Membership-Inference-Angriff tatsächlich ist

Das Leck ist subtiler als „das Modell spuckt deine Krankenakte aus“. Es geht um die *Mitgliedschaft* — die bloße Tatsache, dass deine Daten im Trainingsdatensatz waren.
Beginnen wir damit, was ein solches Modell normalerweise tut. Man gibt ihm etwa ein Thoraxröntgenbild, und es liefert Wahrscheinlichkeiten zurück: 78 % für eine Lungenentzündung, dazu Einschätzungen für Kardiomegalie, Ödem oder Konsolidierung. Genau diese alltägliche diagnostische Antwort ist das *Einzige*, was der Angriff verwendet. Nichts Exotisches.
Die ausgenutzte Schwäche lautet: Ein Modell ist bei genau den Beispielen, auf denen es

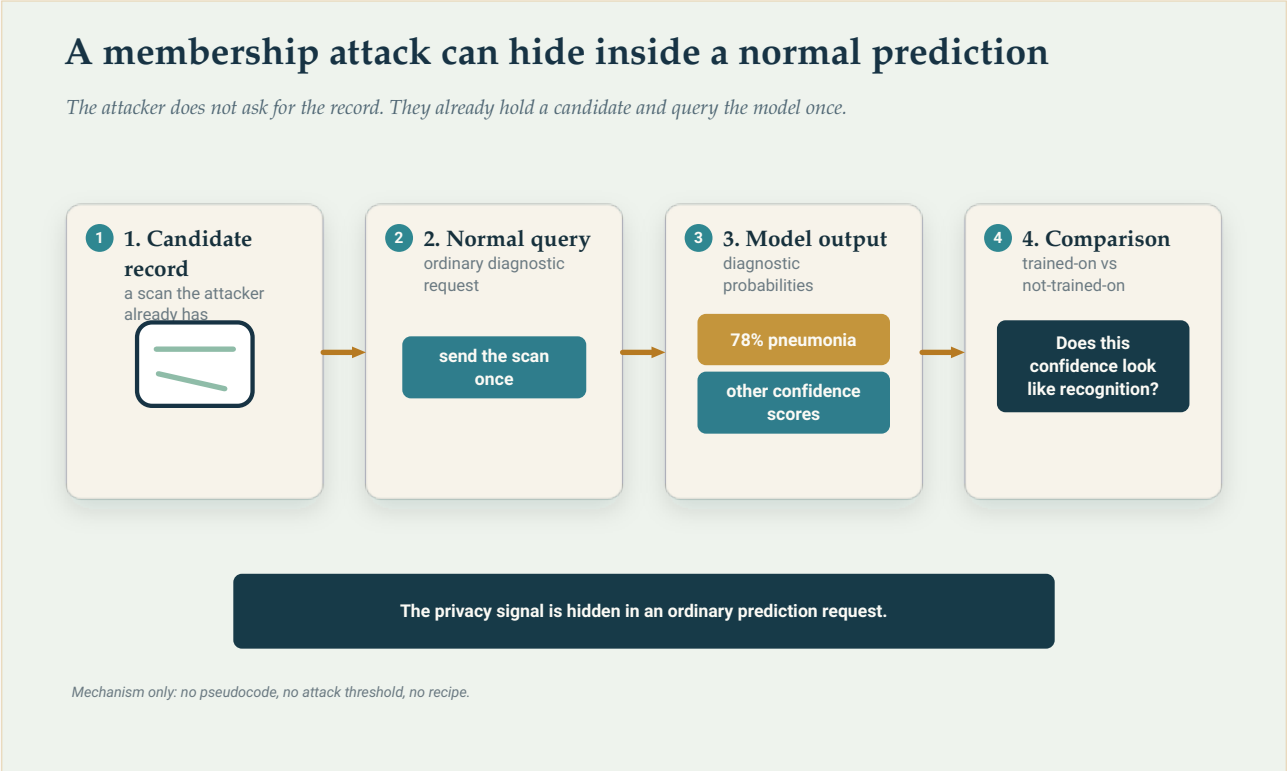
trainiert wurde, häufig ein wenig **sicherer** als bei solchen, die es nie gesehen hat. Man kann sich das wie einen Schüler vorstellen, der die Prüfung heimlich vorher gesehen hat — bei den geübten Fragen kommen die Antworten etwas zu schnell und etwas zu selbstsicher. Aus diesem Hinweis zu erschließen, ob ein bestimmter Datensatz zu den Trainingsbeispielen gehörte, ist das, was „Membership Inference“ bedeutet.

Der Angriff läuft also so ab. Jemand möchte wissen, ob das Bild einer bestimmten Person zum Training des Modells verwendet wurde. Er fragt das Modell **nicht** nach dem Datensatz — ein Kandidatenbild liegt bereits vor, etwa das Bild der Person selbst oder eine sehr ähnliche Kopie. Dieses Bild wird einmal als ganz normale diagnostische Anfrage an das Modell geschickt, und die Sicherheit der Antwort wird notiert. Dann wird geprüft, ob dieses Konfidenzmuster eher einem Modell ähnelt, das auf dem Bild trainiert wurde, oder einem, das es nicht kannte. Diesen Vergleich kann der Angreifer relativ günstig lernen, indem er selbst ein Ersatzmodell („Referenzmodell“) trainiert, um das charakteristische Übermaß an Sicherheit zu erkennen — auf einem normalen Computer und ohne besondere Hardware. Ist das echte Modell bei diesem Bild ungewöhnlich sicher, als würde es es wiedererkennen, ist das ein starkes Indiz dafür, dass das Bild im Trainingsdatensatz lag.

Das Beunruhigende daran: An der Anfrage selbst sieht nichts wie ein Angriff aus. Es ist dieselbe diagnostische Anfrage, die auch ein Arzt stellen würde; das Datenschutzsignal steckt in einer gewöhnlichen Vorhersage. Genau deshalb ist das Risiko konkret — auch wenn es bislang unter klar benannten Laborannahmen demonstriert und nicht als realer Angriff in freier Wildbahn beobachtet wurde.

Warum ist die Mitgliedschaft überhaupt wichtig, wenn der Datensatz selbst nie herauskommt? Weil *Mitgliedschaft eine Information ist*. Wenn ein Modell auf Patienten trainiert wurde, die eine bestimmte Krebsimmuntherapie erhielten, dann verrät die Bestätigung, dass dein Datensatz dazugehört, wahrscheinlich auch, dass du diesen Krebs hattest — genau die Art von Information, die ein Versicherer oder Arbeitgeber niemals erschließen können sollte. Der Inhalt bleibt verschlossen; entweichen kann die Tatsache der Zugehörigkeit.

Die Autoren legen die Annahmen offen dar — Zugriff auf die normalen Vorhersagen des Modells, einen Kandidatendatensatz und ein eigenes Referenzmodell des Angreifers — nicht als Anleitung, sondern damit die Bedrohung konkret analysiert statt vage weggewischt werden kann.



Der Angriff braucht keine spezielle Datenschutz-API. Eine normale diagnostische Anfrage kann ein Signal zur Trainingsmitgliedschaft verraten, wenn das Modell bei einem bereits gesehenen Datensatz ungewöhnlich sicher ist.

Original Aurora diagram — The Clean Paper CC BY 4.0

Was sie herausfanden

Drei Ergebnisse, jedes schärfer als das vorherige.

Durchschnittswerte verbergen die Gefährdeten. Aggregiert gemessen — wie es üblicherweise geschieht — sehen viele dieser Modelle beruhigend privat aus: Für die meisten Patienten ist der Angriff nicht besser als Zufall. Die Sicht auf einzelne Patienten ergab ein anderes Bild. Wie Knolle es in der Mitteilung zur Studie zusammenfasste, hätten frühere Bewertungen nur das durchschnittliche Risiko über alle Patienten gemessen; die neue Analyse untersuche erstmals das Risiko auf Ebene einzelner Patienten und ergebe ein deutlich anderes Bild. Bei manchen Personen gelingt der Angriff **nahezu perfekt** — gemessen als Angriffs-AUC von **0,95 oder höher** (0,5 entspricht einem Münzwurf, 1,0 perfekter Trennung) — obwohl der Durchschnitt über den Datensatz kaum besser als Zufall aussieht.



Der Durchschnitt kann nahe am Zufallsniveau liegen, während ein kleiner rechter Rand der Datensätze deutlich stärker identifizierbar bleibt. Der Punkt der Arbeit: Datenschutz muss auf Patientenebene geprüft werden, nicht nur aggregiert.

Original Aurora diagram — The Clean Paper CC BY 4.0

Die Gefährdung ist ungleich — und trifft die Unterrepräsentierten. Die Patienten, die für einen nahezu perfekten Angriff am anfälligsten waren, gehörten systematisch zu Gruppen, die in den Daten **unterrepräsentiert** waren: ethnische Minderheiten, seltene Krankheitsphänotypen, ungewöhnliche Bildmerkmale. Im Datensatz elektronischer Gesundheitsakten tauchten schwarze Patienten unter den am stärksten gefährdeten Datensätzen **31 % häufiger** auf als erwartet; im Mammografie-Datensatz waren Aufnahmen, die als malignitätsverdächtig markiert waren, in diesem Hochrisikobereich um bemerkenswerte **+1.179 %** überrepräsentiert. (Diese beiden Zahlen sind Beispiele, nicht das gesamte Bild — die Studie berichtet dieselbe Schieflage über mehrere Gruppen hinweg — und es handelt sich um *relative* Überrepräsentationen innerhalb des kleinen extremen Risikorands: also darum, wie viel häufiger diese Patienten unter den am stärksten exponierten Datensätzen auftauchen, nicht um den Anteil aller schwarzen Patienten oder aller verdächtigen Mammografien, die exponiert sind.) Ein Modell hat weniger ähnliche Beispiele, zwischen denen solche Patienten „untergehen“ können; ihre Datensätze stechen deshalb stärker hervor. Und genau dieses Hervorstechen erkennt ein Membership-Inference-Angriff. Das Datenschutzversagen ist nicht zufällig; es konzentriert sich auf Menschen, die ohnehin am Rand stehen.

Größere Modelle machen es schlimmer. Die Zahl der Patienten, die nahezu perfekten Angriffen ausgesetzt waren, stieg stark mit der Modellkapazität. In einem Dermatologie-Datensatz wuchs der Anteil von praktisch null beim kleinsten Modell auf etwa **einen von zehn** beim größten. Während medizinische KI-Modelle immer größer und leistungsfähiger werden — genau die Richtung, in die sich

das Feld bewegt — wird dieses konkrete Risiko nach Warnung der Autoren stärker, nicht schwächer.

Rückerts Zusammenfassung ist deutlich: „Das ist kein tolerierbares Risiko. Gesundheitsdaten sind hochsensibel."

Warum „im Durchschnitt sicher" der falsche Test ist

Es liegt nahe, ein niedriges durchschnittliches Risiko als gesundheitliches Unbedenklichkeitszeugnis für den Datenschutz zu lesen. Die Kernaussage der Arbeit ist, dass der Durchschnitt hier nicht nur ungenau ist — er misst das Falsche.

Eine Datenschutzgarantie ist nur dann sinnvoll, wenn sie auch für die Person mit dem höchsten Risiko gilt, nicht nur für die Person in der Mitte. Ein Modell, bei dem 999 von 1.000 Patienten nicht identifizierbar sind, aber einer nahezu perfekt erkannt werden kann, ist für diesen einen Menschen in keinem sinnvollen Sinn „zu 99,9 % privat". Und wenn dieser eine Mensch vorhersagbar der Patient mit seltener Krankheit oder aus einer ethnischen Minderheit ist, dann ist die Kennzahl nicht bloß unvollständig, sondern still diskriminierend. Sie misst Sicherheit für die Mehrheit und nennt sie Sicherheit für alle.

Genau diese Verschiebung erzwingt die Arbeit: weg von *Wie privat ist dieses Modell im Durchschnitt?* hin zu *Wer ist der am stärksten exponierte Patient — und zu welcher Gruppe gehört er?* Das sind unterschiedliche Fragen. Nur die zweite ist wirklich eine Datenschutzfrage.

Was dies *nicht* beweist — oder behauptet

- Es heißt **nicht**, dass medizinische KI aufgegeben werden sollte. Die Autoren argumentieren für Gegenmaßnahmen, nicht für Rückzug: Risiko messen und reduzieren, nicht den Bau solcher Systeme einstellen.
- Es bedeutet **nicht**, dass jedes medizinische Modell Daten verrät oder dass irgendein konkretes eingesetztes Modell bereits angegriffen wurde. Die Arbeit zeigt, dass das *Risiko existiert und ungleich verteilt ist* und dass übliche aggregierte Kennzahlen es übersehen.
- Es bedeutet **nicht**, dass deine Gesundheitsdaten bereits offengelegt sind. Der Angriff braucht bestimmte Voraussetzungen — Zugriff auf das Modell, einen Kandidatendatensatz und die eigene Infrastruktur des Angreifers — und ist keine beiläufige Fähigkeit.
- Es zeigt **nicht**, dass der *Inhalt* von Patientenakten herausgegeben wird. Was verraten werden kann, ist die Mitgliedschaft — die Tatsache der Aufnahme in den Trainingsdatensatz — und genau diese kann aus einem anderen Grund gefährlich sein.
- Es reduziert die Ungleichheiten **nicht** auf eine einzige Schlagzeilenzahl. Das robuste, wiederkehrende Ergebnis ist das *Muster*: Aggregierte Kennzahlen unterschätzen individuelles Risiko, und unterrepräsentierte Patienten tragen den größten Teil davon — über mehrere Datensätze und Hunderte Modelle hinweg.

Wie belastbar sind die Belege?

Dies ist ein empirisches, methodisches Ergebnis — und ein robustes. Das Muster, dass aggregierte Datenschutzkennzahlen das Risiko einzelner Patienten systematisch unterschätzen, dass das verbleibende Risiko sich auf unterrepräsentierte Gruppen konzentriert und mit der Modellkapazität zunimmt, zeigte sich über **sieben Datensätze mit unterschiedlichen Datentypen** und eine große Zahl von Modellen pro Datensatz. Genau diese Breite macht eine Messbehauptung glaubwürdig und weniger wahrscheinlich zu einem Artefakt eines einzelnen Datensatzes.

Zwei ehrliche Einschränkungen. Erstens demonstriert die Studie ein *Risiko*, indem die von den Autoren entwickelten Angriffe ausgeführt werden. Sie charakterisiert, wie erfolgreich ein fähiger Angreifer unter den angegebenen Annahmen *sein könnte*, nicht wie häufig solche Angriffe in der realen Welt vorkommen. Zweitens beschreiben die eindrucksvollen Zahlen — nahezu perfekte Angriffserfolge bei einzelnen Patienten — absichtlich die am schlechtesten geschützten Individuen. Genau darum geht es. Sie sollten gelesen werden als: „Der Risikorand ist sehr viel schwerer, als der Durchschnitt vermuten lässt“, nicht als: „Die meisten Patienten sind exponiert.“

Angemessen ist weder Alarmismus noch Abwiegeln: Es handelt sich um eine sorgfältige Studie über mehrere Datensätze hinweg, die zeigt, dass die übliche Art, Datenschutz bei medizinischer KI zu zertifizieren, für die eigenen schlimmsten Fälle blind ist — und dass diese schlimmsten Fälle gerade jene Patienten treffen, die ohnehin den geringsten Sicherheitsspielraum haben.

Warum das wichtig ist

Zwei Dinge machen daraus mehr als eine technische Fußnote.

Erstens verändert die Arbeit, was „datenschutzwahrend“ überhaupt bedeuten dürfen sollte. Wird ein Modell mit einem durchschnittlichen Datenschutzwert veröffentlicht, kann dieser Wert tatsächlich niedrig sein und dennoch eine Untergruppe von Patienten verbergen, deren Trainingsmitgliedschaft durch einen Angriff nahezu perfekt erkannt werden kann. Die praktische Forderung der Autoren ist konkret: Datenschutzrisiken **für jeden Patienten** vor der Veröffentlichung bewerten, kontrollieren, wer auf eingesetzte Modelle zugreifen kann, und Verfahren wie **Differential Privacy** einsetzen — kleine, sorgfältig kalibrierte Störungen während des Trainings, die Membership-Inference-Angriffe abschwächen, allerdings mit realen und bewusst zu steuernden Kosten für die Nützlichkeit des Modells. Der Zielkonflikt zwischen Datenschutz und Nutzen ist ausdrücklich vorhanden, nicht kostenlos. „Wir haben den Durchschnitt geprüft“ sollte nicht mehr als ausreichende Prüfung gelten.

Zweitens verknüpft die Arbeit Datenschutz mit Fairness. Dieselben Gruppen, die in medizinischen Daten unterrepräsentiert sind — und deshalb von medizinischer KI ohnehin schlechter bedient werden können — sind zugleich diejenigen, deren Privatsphäre diese KI am schlechtesten schützt. Ein Forschungsfeld, das intensiv an der ersten Ungleichheit arbeitet, kann die zweite nicht als Problem einer anderen Abteilung behandeln. Es sind dieselben Menschen.

Die beruhigende Erzählung zum Datenschutz medizinischer KI — *wir haben gemessen, der Durchschnitt*

ist niedrig, also ist alles in Ordnung — wird von dieser Arbeit zerlegt. Nicht, um Menschen von der Technologie abzuschrecken, sondern um den Maßstab dorthin zu verschieben, wo er hingehört: zu einem Versprechen, das man nur dann als eingehalten bezeichnen kann, wenn man es für die Person geprüft hat, die am ehesten Schaden nehmen könnte.

Kurz zusammengefasst

Membership-Inference-Angriffe versuchen herauszufinden, ob der Datensatz einer bestimmten Person zum Training eines KI-Modells gehörte. Weil diese Zugehörigkeit selbst sensible Informationen verraten kann — etwa dass jemand eine bestimmte Krankheit hatte —, ist das ein echtes Datenschutzleck, selbst wenn der zugrunde liegende Datensatz nie sichtbar wird. Frühere Arbeiten maßen, wie oft solche Angriffe *im Durchschnitt* über einen Datensatz erfolgreich sind. Diese Studie maß das Risiko *pro Patient* über sieben medizinische Datensätze hinweg — Bildgebung, EKG und elektronische Gesundheitsakten — und über viele Modelle pro Datensatz. Das Ergebnis: Aggregierte Kennzahlen unterschätzen das Risiko stark. Einzelne Personen können nahezu perfekt identifiziert werden (Angriffs-AUC $\geq$ 0,95), selbst wenn der Durchschnitt über den Datensatz sicher aussieht; am stärksten gefährdet sind systematisch Menschen aus unterrepräsentierten Gruppen — ethnische Minderheiten, seltene Erkrankungen, ungewöhnliche Bildmerkmale — und das Problem wächst mit der Modellgröße. Die Autoren fordern nicht, medizinische KI aufzugeben, sondern Datenschutz individuell zu messen, den Modellzugriff zu kontrollieren und Differential Privacy einzusetzen. Die Kernaussage: „im Durchschnitt privat“ ist keine Datenschutzgarantie, und versagen kann sie gerade bei den Menschen, die ohnehin am wenigsten geschützt sind.

Nüchtern geprüft

Was die Arbeit zeigt: Über sieben medizinische Datensätze und viele Modelle pro Datensatz hinweg ist das patientenspezifische Risiko durch Membership Inference bei manchen Personen sehr viel höher, als aggregierte Kennzahlen vermuten lassen — bis hin zu nahezu perfektem Angriffserfolg (AUC $\geq$ 0,95). Dieses verbleibende Risiko trifft überproportional unterrepräsentierte Gruppen und wächst mit der Modellkapazität.

Was plausibel, aber nicht bewiesen ist: Dass reale Angreifer dies bereits gegen eingesetzte klinische Modelle ausnutzen. Die Studie demonstriert die *Fähigkeit und ihre Verteilung* unter definierten Annahmen, nicht die Häufigkeit tatsächlicher Angriffe.

Was die Arbeit nicht zeigt: Dass medizinische KI aufgegeben werden sollte; dass jedes Modell leckt oder irgendein konkretes eingesetztes Modell kompromittiert wurde; dass die *Inhalte* von Patientenakten statt nur die Trainingsmitgliedschaft offengelegt werden; oder eine einzige Zahl, die die gesamte Ungleichheit beschreibt — robust ist das Muster, nicht eine einzelne Kennzahl.

Wichtigste Einschränkungen: Gemessen wird ein Worst-Case-Risiko mit den eigenen Angriffen der Autoren, nicht die Zahl beobachteter Vorfälle; die dramatischen Werte beschreiben bewusst die am

stärksten exponierten Personen; und die genauen gruppenspezifischen Unterschiede erscheinen als konsistentes Muster über Datensätze hinweg, nicht als eine einzige zusammenfassende Zahl.

Wie viel Vertrauen sollte ein allgemeiner Leser haben? Hohes Vertrauen darin, dass aggregierte Datenschutzkennzahlen individuelles Risiko unterschätzen, dass das verbleibende Risiko sich auf unterrepräsentierte Patienten konzentriert und dass es mit der Modellgröße zunimmt — das wird über viele Datensätze und Modelle demonstriert. Mittleres Vertrauen hinsichtlich der realen Häufigkeit solcher Angriffe, die diese Studie nicht misst. Die sichere Lesart lautet weder „medizinische KI verrät deine Daten“ noch „Datenschutz ist gelöst“, sondern: „Die übliche Datenschutzprüfung ist für ihre eigenen schlimmsten Fälle blind, und diese Fälle treffen die verletzlichsten Patienten — also muss sich die Prüfung ändern.“

Quellen

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