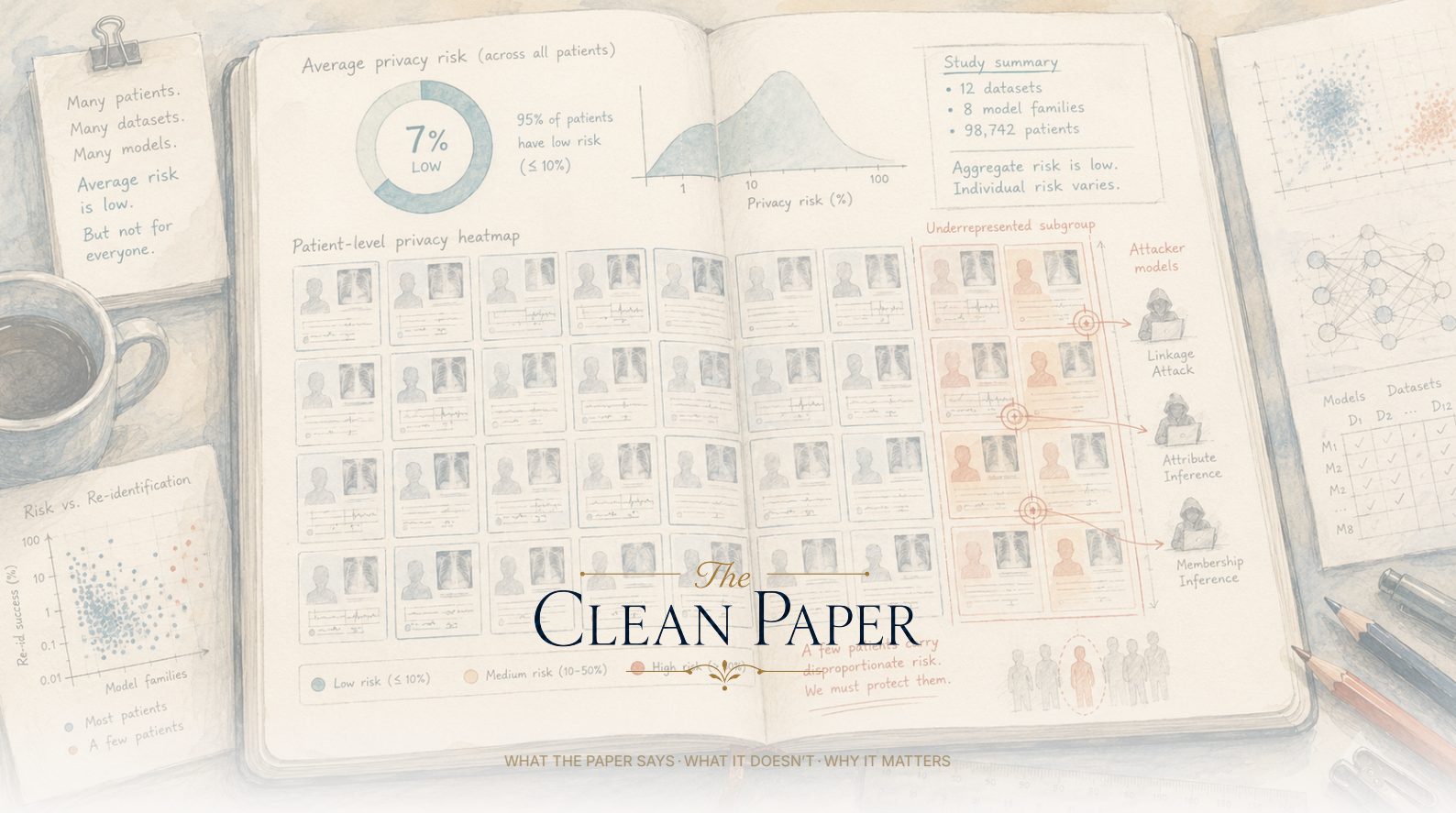


Μια ιατρική ΤΝ μπορεί να είναι ιδιωτική κατά μέσο όρο και παρ' όλα αυτά να εκθέτει συγκεκριμένους ασθενείς — κυρίως τους υποεκπροσωπούμενους

Ένα μοντέλο ιατρικής ΤΝ που χαρακτηρίζεται «προστατευτικό της ιδιωτικότητας» συνήθως στηρίζεται σε έναν μέσο αριθμό. Αυτή η μελέτη υποστηρίζει ότι αυτός είναι λάθος έλεγχος. Μελετώντας επιθέσεις membership inference — οι οποίες αποκαλύπτουν αν η εγγραφή ενός συγκεκριμένου ανθρώπου βρισκόταν στα δεδομένα εκπαίδευσης ενός μοντέλου και έτσι μπορούν να προδώσουν ότι είχε μια συγκεκριμένη νόσο — οι συγγραφείς μέτρησαν τον κίνδυνο ανά ασθενή αντί συνολικά, σε επτά ιατρικά σύνολα δεδομένων (απεικόνιση, ΗΚΓ, ηλεκτρονικοί φάκελοι) και πολλά μοντέλα για το καθένα. Το μοτίβο: μοντέλα που φαίνονται ασφαλή κατά μέσο όρο μπορούν ακόμη να επιτρέπουν σε έναν επιτιθέμενο να ταυτοποιεί συγκεκριμένα άτομα σχεδόν τέλεια (AUC επίθεσης ≥ 0.95)· όσοι εκτίθενται ανήκουν συστηματικά σε υποεκπροσωπούμενες ομάδες (εθνοτικές μειονότητες, σπάνια νόσος, ασυνήθιστη απεικόνιση)· και το πρόβλημα χειροτερεύει όσο μεγαλώνουν τα μοντέλα. Οι συγγραφείς δεν λένε να εγκαταλείψουμε την ιατρική ΤΝ — λένε να μετράμε την ιδιωτικότητα ανά ασθενή, να ελέγχουμε την πρόσβαση στα μοντέλα και να χρησιμοποιούμε differential privacy. Ο άβολος πυρήνας: «ιδιωτικό κατά μέσο όρο» δεν είναι εγγύηση ιδιωτικότητας, και αποτυγχάνει ακριβώς για τους ασθενείς που ήδη προστατεύονται λιγότερο.

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

Μια εγγύηση ιδιωτικότητας είναι υπόσχεση προς έναν άνθρωπο, όχι προς έναν μέσο όρο

Όταν ένα μοντέλο ιατρικής ΤΝ χαρακτηρίζεται «προστατευτικό της ιδιωτικότητας», ο ισχυρισμός αυτός συνήθως στηρίζεται σε έναν αριθμό: στο σύνολο των ασθενών των οποίων τα δεδομένα χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου, η μέση πιθανότητα να συναχθεί αν ένα συγκεκριμένο άτομο ανήκε στο σύνολο εκπαίδευσης είναι χαμηλή. Ακούγεται καθησυχαστικό. Το ήσυχο, άβολο συμπέρασμα αυτής της εργασίας είναι ότι πρόκειται για λάθος αριθμό.

Η ιδιωτικότητα δεν είναι μέσος όρος. Είναι μια υπόσχεση προς κάθε άνθρωπο ξεχωριστά — ότι η συμμετοχή του στο σύνολο εκπαίδευσης δεν θα επιστρέψει αργότερα για να τον εκθέσει. Και ένας μέσος όρος μπορεί να τηρεί αυτή την υπόσχεση για σχεδόν όλους, ενώ την παραβιάζει πλήρως για λίγους. Οι ερευνητές μέτρησαν τον κίνδυνο ιδιωτικότητας έναν ασθενή τη φορά, με ανάλυση που δεν είχε χρησιμοποιηθεί προηγουμένως, και βρήκαν ακριβώς αυτό: μοντέλα που συνολικά φαίνονται ασφαλή μπορούν να αποκαλύπτουν σχεδόν τέλεια τη συμμετοχή συγκεκριμένων ανθρώπων — και όσοι εκτίθενται δυσανάλογα είναι συχνά εκείνοι που ήδη προστατεύονται λιγότερο.

Τι έκαναν οι συγγραφείς

Η ομάδα — με επικεφαλής τον Moritz Knolle, μαζί με τον Daniel Rückert (και οι δύο στο Technical University of Munich) και τον Georg Kaissis (Hasso Plattner Institute), καθώς και συνεργάτες από το Imperial College London — πήρε μια γνωστή απειλή ιδιωτικότητας που ονομάζεται **επίθεση συμπερασμού συμμετοχής (membership inference attack)** και άλλαξε την ερώτηση. Αντί να ρωτήσει «κατά μέσο όρο, πόσο συχνά πετυχαίνει αυτή η επίθεση σε όλο το σύνολο δεδομένων;», ρώτησε «για αυτόν τον συγκεκριμένο ασθενή, πόσο εκτεθειμένος είναι;»

Έκαναν αυτή την ανάλυση ανά ασθενή σε **επτά καθιερωμένα ιατρικά σύνολα δεδομένων**, που κάλυπταν πολύ διαφορετικούς τύπους δεδομένων — ιατρικές εικόνες, ηλεκτροκαρδιογραφήματα και ηλεκτρονικούς φακέλους υγείας — και, συνολικά, 200 μοντέλα για το καθένα. Για κάθε ασθενή σε κάθε μοντέλο, εκτίμησαν με πόση βεβαιότητα ένας επιτιθέμενος θα μπορούσε να καταλάβει αν η εγγραφή αυτού του ατόμου είχε χρησιμοποιηθεί στην εκπαίδευση και στη συνέχεια ανέλυσαν τα αποτελέσματα ανά ομάδα: κατάσταση νόσου, αυτοδηλωμένη φυλή, ασφάλιση, φύλο και πρωτόκολλο απεικόνισης.

> **Τι είναι στην πράξη μια επίθεση συμπερασμού συμμετοχής**
>
> Η διαρροή εδώ είναι πιο λεπτή από το «το μοντέλο βγάζει τον ιατρικό σου φάκελο». Αφορά τη *συμμετοχή* — και μόνο το γεγονός ότι τα δεδομένα σου βρίσκονταν στο σύνολο εκπαίδευσης.

Ας ξεκινήσουμε από το τι κάνει κανονικά ένα τέτοιο μοντέλο. Του δίνεις μια εξέταση — ας πούμε μια ακτινογραφία θώρακος — και επιστρέφει πιθανότητες: 78% πιθανότητα πνευμονίας, μαζί με εκτιμήσεις για καρδιομεγαλία, οίδημα, πύκνωση. Αυτή η συνηθισμένη διαγνωστική απάντηση είναι το *μόνο* πράγμα που χρησιμοποιεί η επίθεση. Τίποτε εξωτικό.

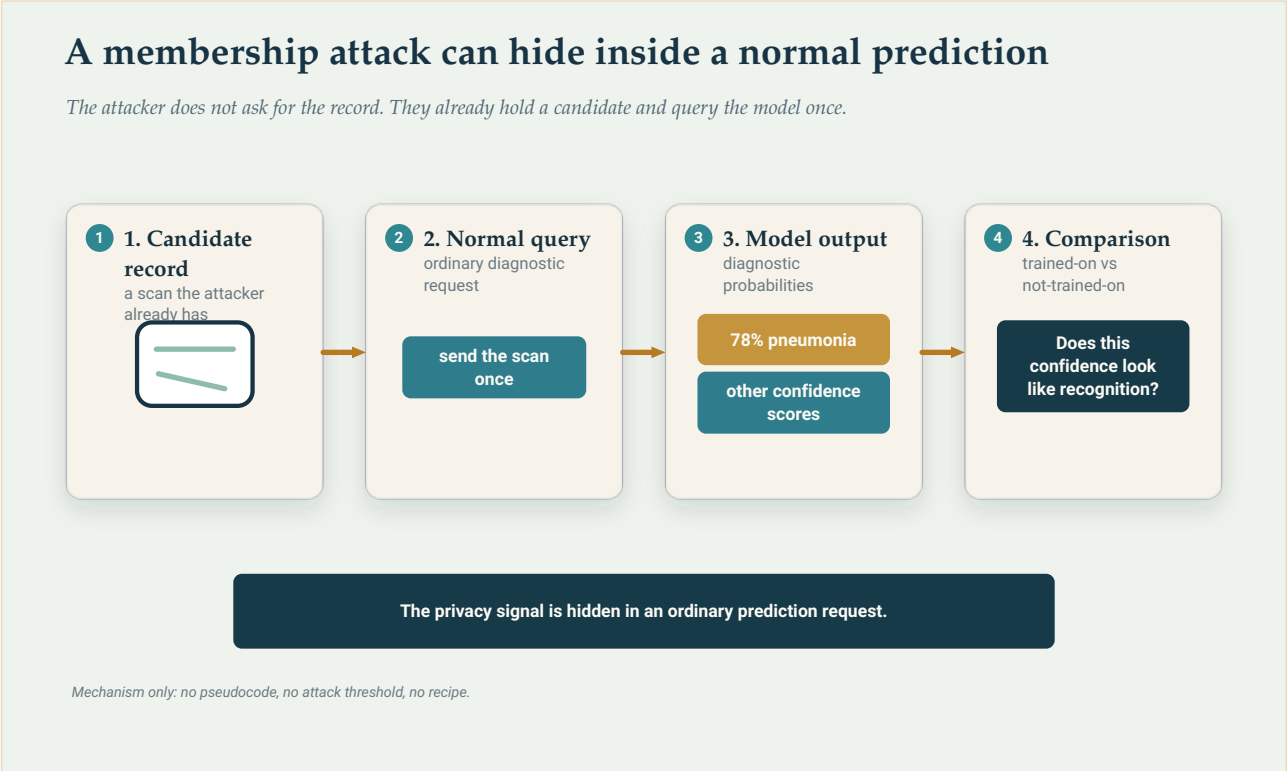
Η αδυναμία στην οποία βασίζεται είναι η εξής: ένα μοντέλο είναι συνήθως λίγο **πιο βέβαιο** για τα ακριβή παραδείγματα πάνω στα οποία εκπαιδεύτηκε απ' ό,τι για παραδείγματα που δεν έχει δει ποτέ. Σκεφτείτε έναν μαθητή που είχε κρυφά δει το διαγώνισμα από πριν — στις ερωτήσεις που είχε ήδη εξασκηθεί, οι απαντήσεις έρχονται λίγο πιο γρήγορα, λίγο πιο σίγουρα. Το να συμπεράνει κανείς από αυτή την ένδειξη αν μια συγκεκριμένη εγγραφή ήταν ένα από τα παραδείγματα που «μελέτησε» το μοντέλο είναι αυτό που σημαίνει «membership inference».

Η επίθεση λοιπόν λειτουργεί κάπως έτσι, βήμα προς βήμα. Κάποιος θέλει να μάθει αν η εξέταση ενός συγκεκριμένου ανθρώπου χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου. **Δεν** ζητά από το μοντέλο τον φάκελο — έχει ήδη μια υποψήφια εξέταση (την εξέταση του ίδιου του ατόμου ή ένα πολύ κοντινό αντίγραφο). Τη στέλνει μία φορά, ως μια συνηθισμένη διαγνωστική ερώτηση, και σημειώνει πόσο βέβαιη είναι η απάντηση. Μετά ελέγχει αν αυτή η βεβαιότητα μοιάζει περισσότερο με μοντέλο που *είχε* εκπαιδευτεί πάνω στην εξέταση ή με μοντέλο που *δεν είχε* — σύγκριση που μπορεί να κάνει σχετικά φθηνά εκπαιδεύοντας ένα δικό του υποκατάστατο μοντέλο (ένα «μοντέλο αναφοράς») ώστε να μάθει πώς μοιάζει αυτή η χαρακτηριστική υπερβολική αυτοπεποίθηση, σε έναν συνηθισμένο υπολογιστή χωρίς ειδικό υλικό. Αν το πραγματικό μοντέλο είναι ασυνήθιστα βέβαιο γι' αυτή την εξέταση — σαν να την αναγνωρίζει — αυτό αποτελεί ισχυρή ένδειξη ότι η εξέταση βρισκόταν στα δεδομένα εκπαίδευσης.

Το ανησυχητικό είναι ότι τίποτε στο αίτημα δεν μοιάζει με επίθεση: είναι η ίδια διαγνωστική ερώτηση που θα έκανε ένας κλινικός γιατρός και το σήμα ιδιωτικότητας κρύβεται μέσα σε μια συνηθισμένη πρόβλεψη. Γι' αυτό ακριβώς ο κίνδυνος είναι συγκεκριμένος — παρότι μέχρι στιγμής έχει αποδειχθεί υπό δηλωμένες εργαστηριακές παραδοχές και όχι σε καταγεγραμμένες επιθέσεις στον πραγματικό κόσμο.

Γιατί έχει σημασία η συμμετοχή, αν ο ίδιος ο φάκελος δεν διαρρέει ποτέ; Επειδή *η συμμετοχή είναι πληροφορία*. Αν ένα μοντέλο εκπαιδεύτηκε σε ασθενείς που έλαβαν μια συγκεκριμένη ανοσοθεραπεία για καρκίνο, η επιβεβαίωση ότι η δική σου εγγραφή περιλαμβανόταν εκεί αποκαλύπτει ότι πιθανότατα είχες αυτόν τον καρκίνο — ακριβώς το είδος πληροφορίας που ένας ασφαλιστής ή εργοδότης δεν θα έπρεπε ποτέ να μπορεί να συμπεράνει. Το περιεχόμενο παραμένει κλειστό· αυτό που διαφεύγει είναι το γεγονός της συμμετοχής.

Οι συγγραφείς δηλώνουν καθαρά αυτές τις προϋποθέσεις — πρόσβαση στις συνηθισμένες προβλέψεις του μοντέλου, μια υποψήφια εγγραφή και ένα μοντέλο αναφοράς του ίδιου του επιτιθέμενου — όχι ως οδηγίες επίθεσης, αλλά για να μπορεί ο κίνδυνος να αναλυθεί συγκεκριμένα αντί να αντιμετωπίζεται αόριστα.



Η επίθεση δεν χρειάζεται κάποιο ειδικό API ιδιωτικότητας. Μια συνηθισμένη διαγνωστική ερώτηση μπορεί να διαρρεύσει σήμα συμμετοχής αν το μοντέλο είναι ασυνήθιστα βέβαιο για μια εγγραφή που έχει ξαναδεί.

Τι βρήκαν

Τρία ευρήματα, το καθένα πιο αιχμηρό από το προηγούμενο.

Οι μέσοι όροι κρύβουν τους εκτεθειμένους. Όταν μετρώνται συνολικά — όπως γίνεται συνήθως — πολλά από αυτά τα μοντέλα φαίνονται καθησυχαστικά ιδιωτικά: για τους περισσότερους ασθενείς η επίθεση δεν τα καταφέρνει καλύτερα από την τύχη. Η εικόνα ανά ασθενή ήταν διαφορετική. Όπως το έθεσε ο Knolle στην ανακοίνωση της μελέτης, οι προηγούμενες αξιολογήσεις «μετρούσαν μόνο τον μέσο κίνδυνο σε όλους τους ασθενείς. Εξετάσαμε για πρώτη φορά τον κίνδυνο στο επίπεδο του κάθε ασθενούς — και η εικόνα που προκύπτει είναι πολύ διαφορετική». Για ορισμένα άτομα η επίθεση πετυχαίνει **σχεδόν τέλεια** — οι συγγραφείς το μετρούν ως AUC επίθεσης **0.95 ή μεγαλύτερο** (0.5 αντιστοιχεί σε κορώνα-γράμματα, 1.0 σε τέλεια διάκριση) — ακόμη κι όταν ο μέσος όρος σε όλο το σύνολο δεδομένων δεν ήταν καλύτερος από την τύχη.

Ο μέσος όρος μπορεί να βρίσκεται κοντά στο επίπεδο της τύχης ενώ μια μικρή «ουρά» εγγραφών παραμένει πολύ περισσότερο εκτεθειμένη. Το βασικό σημείο της εργασίας είναι ότι η ιδιωτικότητα πρέπει να ελέγχεται στο επίπεδο του ασθενούς, όχι μόνο συνολικά.

Η έκθεση δεν είναι ίση — και πλήττει τους υποεκπροσωπούμενους. Οι ασθενείς που ήταν πιο ευάλωτοι σε μια σχεδόν τέλεια επίθεση ανήκαν συστηματικά σε ομάδες **υποεκπροσωπούμενες** στα **δεδομένα**: εθνοτικές μειονότητες, σπάνιοι φαινότυποι νόσων, ασυνήθιστα χαρακτηριστικά απεικόνισης. Στο σύνολο ηλεκτρονικών φακέλων, οι μαύροι ασθενείς εμφανίζονταν **31%** **συχνότερα** από το αναμενόμενο ανάμεσα στις πιο ευάλωτες εγγραφές· στο σύνολο μαστογραφιών, οι εξετάσεις που είχαν χαρακτηριστεί ύποπτες για κακοήθεια ήταν υπερεκπροσωπημένες σε αυτή τη ζώνη κινδύνου κατά εντυπωσιακό **+1,179%**. (Αυτοί οι δύο αριθμοί είναι παραδείγματα, όχι ολόκληρη η εικόνα — η εργασία αναφέρει την ίδια στρέβλωση σε αρκετές ομάδες — και πρόκειται για *σχετική* υπερεκπροσώπηση μέσα στη μικρή ουρά ακραίου κινδύνου: πόσο συχνότερα εμφανίζονται αυτοί οι ασθενείς ανάμεσα στις πιο εκτεθειμένες εγγραφές, όχι ποιο ποσοστό όλων των μαύρων ασθενών ή όλων των ύποπτων μαστογραφιών είναι εκτεθειμένο.) Ένα μοντέλο έχει λιγότερα παρόμοια παραδείγματα μέσα στα οποία μπορούν να «χαθούν» αυτοί οι ασθενείς, οπότε οι εγγραφές τους ξεχωρίζουν — και ακριβώς αυτό το ξεχώρισμα ανιχνεύει μια επίθεση membership inference. Η αποτυχία ιδιωτικότητας δεν είναι τυχαία· συγκεντρώνεται στους ανθρώπους που βρίσκονται ήδη στο περιθώριο.

Τα μεγαλύτερα μοντέλα το κάνουν χειρότερο. Ο αριθμός των ασθενών που εκτέθηκαν σε σχεδόν τέλειες επιθέσεις αυξήθηκε απότομα μαζί με τη χωρητικότητα του μοντέλου — σε ένα σύνολο δεδομένων δερματολογίας, το ποσοστό ανέβηκε από ουσιαστικά μηδέν στο μικρότερο μοντέλο σε περίπου **έναν στους δέκα** στο μεγαλύτερο. Καθώς τα μοντέλα ιατρικής ΤΝ γίνονται μεγαλύτερα και ικανότερα — προς την κατεύθυνση όπου κινείται

ολόκληρος ο κλάδος — αυτός ο συγκεκριμένος κίνδυνος, προειδοποιούν οι συγγραφείς, γίνεται σοβαρότερος, όχι μικρότερος.

Η σύνοψη του Rückert είναι ωμή: «Αυτός δεν είναι ανεκτός κίνδυνος. Τα δεδομένα υγείας είναι εξαιρετικά ευαίσθητα.»

Γιατί το «ασφαλές κατά μέσο όρο» είναι λάθος τεστ

Είναι δελεαστικό να διαβάζουμε έναν χαμηλό μέσο κίνδυνο σαν καθαρό πιστοποιητικό υγείας. Το κεντρικό μάθημα της εργασίας είναι ότι η χρήση του μέσου όρου εδώ δεν είναι απλώς ανακριβής — μετρά το λάθος πράγμα.

Μια εγγύηση ιδιωτικότητας έχει νόημα μόνο αν ισχύει για τον άνθρωπο που κινδυνεύει περισσότερο, όχι για τον άνθρωπο στη μέση της κατανομής. Ένα μοντέλο όπου 999 από 1.000 ασθενείς δεν μπορούν να αναγνωριστούν αλλά ένας μπορεί να ταυτοποιηθεί σχεδόν τέλεια δεν είναι «99,9% ιδιωτικό» με κανέναν τρόπο που να έχει σημασία γι' αυτόν τον έναν άνθρωπο — και αν αυτός ο ένας είναι προβλέψιμα ο ασθενής με σπάνια νόσο ή ο ασθενής από εθνοτική μειονότητα, η μετρική δεν είναι απλώς ελλιπής· είναι σιωπηρά μεροληπτική. Αναφέρει ασφάλεια για την πλειονότητα και την ονομάζει ασφάλεια για όλους.

Αυτή είναι η μετατόπιση που επιβάλλει η εργασία: από το *πόσο ιδιωτικό είναι αυτό το μοντέλο κατά μέσο όρο;* στο *ποιος είναι ο πιο εκτεθειμένος ασθενής και ποιος είναι αυτός ο άνθρωπος;* Είναι διαφορετικές ερωτήσεις, και μόνο η δεύτερη είναι πραγματικά ερώτηση ιδιωτικότητας.

Τι *δεν* αποδεικνύει — ούτε ισχυρίζεται

- **Δεν** λέει ότι πρέπει να εγκαταλειφθεί η ιατρική ΤΝ. Η προσέγγιση των συγγραφέων είναι μετριασμός του κινδύνου, όχι υποχώρηση: μέτρησε και διόρθωσε τον κίνδυνο, μην σταματήσεις να κατασκευάζεις συστήματα.
- **Δεν** σημαίνει ότι κάθε ιατρικό μοντέλο διαρρέει πληροφορίες ή ότι κάποιο συγκεκριμένο μοντέλο σε χρήση έχει ήδη δεχθεί επίθεση. Δείχνει ότι *ο κίνδυνος υπάρχει και κατανέμεται άνισα* και ότι οι συνηθισμένες συγκεντρωτικές μετρικές τον χάνουν.
- **Δεν** σημαίνει ότι οι δικοί σου φάκελοι είναι ήδη εκτεθειμένοι. Η επίθεση χρειάζεται συγκεκριμένες συνθήκες — πρόσβαση στο μοντέλο, μια υποψήφια εγγραφή και την υποδομή του επιτιθέμενου — δεν είναι μια τυχαία ή άμεση δυνατότητα.
- **Δεν** δείχνει ότι διαρρέει το *περιεχόμενο* του φακέλου του ασθενούς. Αυτό που διαρρέει είναι η συμμετοχή — το γεγονός της ένταξης στο σύνολο εκπαίδευσης — που είναι επικίνδυνο για διαφορετικό λόγο, όχι επειδή το μοντέλο «ξεφορτώνει» τον φάκελο.
- **Δεν** συμπυκνώνει τις ανισότητες σε έναν μόνο εντυπωσιακό αριθμό. Το ισχυρό, αναπαραγώγιμο αποτέλεσμα είναι το μοτίβο — οι συγκεντρωτικές μετρικές υποτιμούν τον ατομικό κίνδυνο

και οι υποεκπροσωπούμενοι ασθενείς επωμίζονται το μεγαλύτερο μέρος του — σε πολλά σύνολα δεδομένων και εκατοντάδες μοντέλα.

Πόσο ισχυρά είναι τα στοιχεία;

Πρόκειται για εμπειρικό, μεθοδολογικό αποτέλεσμα, και μάλιστα ισχυρό. Το μοτίβο — οι συγκεντρωτικές μετρικές ιδιωτικότητας υποτιμούν συστηματικά τον κίνδυνο ανά ασθενή, ο υπολειπόμενος κίνδυνος συγκεντρώνεται σε υποεκπροσωπούμενες ομάδες και επιδεινώνεται με τη χωρητικότητα του μοντέλου — εμφανίστηκε σε **επτά σύνολα δεδομένων διαφορετικών τύπων** και σε μεγάλο αριθμό μοντέλων ανά σύνολο. Αυτή ακριβώς η έκταση κάνει έναν ισχυρισμό μέτρησης πειστικό και όχι τεχνούργημα ενός μόνο dataset.

Δύο έντιμες επιφυλάξεις. Πρώτον, πρόκειται για επίδειξη *κινδύνου*, που μετρήθηκε εκτελώντας τις επιθέσεις που κατασκεύασαν οι ίδιοι οι συγγραφείς· χαρακτηρίζει το πόσο καλά *θα μπορούσε να τα* καταφέρει ένας ικανός επιτιθέμενος υπό τις δηλωμένες παραδοχές, όχι το πόσο συχνά συμβαίνουν τέτοιες επιθέσεις στον πραγματικό κόσμο. Δεύτερον, οι εντυπωσιακοί αριθμοί — σχεδόν τέλεια επιτυχία επίθεσης για ορισμένους ασθενείς — περιγράφουν τους πιο εκτεθειμένους ανθρώπους *εκ σχεδιασμού*. Αυτό ακριβώς είναι το νόημα, και πρέπει να διαβαστούν ως «η ουρά είναι πολύ βαρύτερη απ' όσο υπονοεί ο μέσος όρος», όχι ως «οι περισσότεροι ασθενείς είναι εκτεθειμένοι».

Η κατάλληλη στάση δεν είναι ούτε πανικός ούτε απόρριψη: πρόκειται για μια προσεκτική μελέτη πολλών datasets που δείχνει ότι ο συνήθης τρόπος πιστοποίησης της ιδιωτικότητας στην ιατρική ΤΝ είναι τυφλός στις χειρότερες περιπτώσεις του — και ότι αυτές οι χειρότερες περιπτώσεις αφορούν τους ασθενείς με το μικρότερο περιθώριο προστασίας.

Γιατί έχει σημασία

Δύο πράγματα κάνουν αυτό το αποτέλεσμα περισσότερο από μια τεχνική υποσημείωση.

Πρώτον, αλλάζει το τι θα έπρεπε να επιτρέπεται να σημαίνει «προστατεύει την ιδιωτικότητα». Αν ένα μοντέλο κυκλοφορήσει με έναν μέσο δείκτη ιδιωτικότητας, ο δείκτης αυτός μπορεί πράγματι να είναι χαμηλός και παρ' όλα αυτά να κρύβει μια υποομάδα ασθενών που μπορούν να ταυτοποιηθούν σχεδόν τέλεια μέσω membership inference. Η πρακτική απαίτηση της εργασίας είναι συγκεκριμένη: αξιολόγηση του κινδύνου ιδιωτικότητας **ανά ασθενή** πριν από την κυκλοφορία, έλεγχος του ποιος έχει πρόσβαση στα μοντέλα που αναπτύσσονται και χρήση τεχνικών όπως η **διαφορική ιδιωτικότητα (differential privacy)** — μικρός, προσεκτικά βαθμονομημένος θόρυβος που προστίθεται κατά την εκπαίδευση και αποδυναμώνει τις επιθέσεις membership inference, με πραγματικό και διαχειρίσιμο κόστος στη χρησιμότητα του μοντέλου (ο συμβιβασμός ιδιωτικότητας–χρησιμότητας είναι ρητός, δεν είναι δωρεάν). Το «ελέγξαμε τον μέσο όρο» δεν θα έπρεπε πια να μετρά ως ολοκληρωμένος έλεγχος.

Δεύτερον, συνδέει την ιδιωτικότητα με τη δικαιοσύνη. Οι ίδιες ομάδες που υποεκπροσωπούνται στα ιατρικά δεδομένα — και άρα ήδη εξυπηρετούνται χειρότερα από την ιατρική ΤΝ — αποδεικνύεται ότι είναι και εκείνες των οποίων την ιδιωτικότητα η ίδια ΤΝ προστατεύει λιγότερο. Ένας κλάδος που εργάζεται σοβαρά πάνω στην πρώτη ανισότητα δεν μπορεί να αντιμετωπίζει τη δεύτερη σαν αρμοδιότητα κάποιου άλλου. Πρόκειται για τους ίδιους ανθρώπους.

Η καθησυχαστική ιστορία για την ιδιωτικότητα της ιατρικής ΤΝ — *τη μετρήσαμε, ο μέσος όρος είναι χαμηλός, είμαστε εντάξει* — είναι ακριβώς η ιστορία που αυτή η εργασία αποδομεί. Όχι για να αποθαρρύνει κανέναν από την τεχνολογία, αλλά για να μετακινήσει το πρότυπο εκεί όπου ανήκει: σε μια υπόσχεση που μπορείς να ισχυριστείς ότι τηρείς μόνο αν την έχεις ελέγξει για τον άνθρωπο που είναι πιθανότερο να πληγεί.

Καθαρή σύνοψη

Οι επιθέσεις membership inference προσπαθούν να προσδιορίσουν αν η εγγραφή ενός συγκεκριμένου ανθρώπου βρισκόταν στα δεδομένα εκπαίδευσης ενός μοντέλου ΤΝ — και επειδή η ίδια η συμμετοχή μπορεί να αποκαλύπτει κάτι ευαίσθητο (ότι είχες μια συγκεκριμένη νόσο, για παράδειγμα), αυτό αποτελεί πραγματική διαρροή ιδιωτικότητας ακόμη κι όταν ο υποκείμενος φάκελος δεν εμφανίζεται ποτέ. Προηγούμενες εργασίες μετρούσαν πόσο συχνά πετυχαίνουν τέτοιες επιθέσεις *κατά μέσο όρο* σε ένα dataset. Αυτή η μελέτη το μέτρησε *ανά ασθενή*, σε επτά ιατρικά σύνολα δεδομένων (απεικόνιση, ΗΚΓ, ηλεκτρονικοί φάκελοι) και πολλά μοντέλα για το καθένα, και βρήκε ότι οι συγκεντρωτικές μετρικές υποτιμούν σοβαρά τον κίνδυνο: ορισμένα άτομα μπορούν να ταυτοποιηθούν σχεδόν τέλεια (AUC επίθεσης $\geq$ 0.95) ακόμη κι όταν ο μέσος όρος του dataset φαίνεται ασφαλής· οι πιο ευάλωτοι είναι συστηματικά όσοι ανήκουν σε υποεκπροσωπούμενες ομάδες (εθνοτικές μειονότητες, σπάνιες νόσοι, ασυνήθιστη απεικόνιση)· και το πρόβλημα μεγαλώνει μαζί με το μέγεθος του μοντέλου. Οι συγγραφείς δεν υποστηρίζουν ότι πρέπει να εγκαταλείψουμε την ιατρική ΤΝ — υποστηρίζουν ότι πρέπει να μετράμε την ιδιωτικότητα σε ατομικό επίπεδο, να ελέγχουμε την πρόσβαση στα μοντέλα και να χρησιμοποιούμε differential privacy. Το συμπέρασμα: «ιδιωτικό κατά μέσο όρο» δεν είναι εγγύηση ιδιωτικότητας, και εκείνοι στους οποίους αποτυγχάνει είναι συνήθως όσοι ήδη προστατεύονται λιγότερο.

Έλεγχος χωρίς υπερβολές

Τι δείχνει η εργασία: Σε επτά ιατρικά σύνολα δεδομένων και πολλά μοντέλα για το καθένα, ο κίνδυνος membership inference ανά ασθενή είναι για ορισμένα άτομα πολύ υψηλότερος απ’ όσο υπονοούν οι συγκεντρωτικές μετρικές — έως σχεδόν τέλεια επιτυχία επίθεσης (AUC $\geq$ 0.95) — και αυτός ο υπολειπόμενος κίνδυνος πλήττει δυσανάλογα υποεκπροσωπούμενες ομάδες και αυξάνεται μαζί με τη χωρητικότητα του μοντέλου.

Τι είναι εύλογο αλλά δεν έχει αποδειχθεί: Ότι επιτιθέμενοι στον πραγματικό κόσμο ήδη εκμεταλλεύονται αυτή την αδυναμία σε κλινικά μοντέλα που βρίσκονται σε χρήση· η μελέτη δείχνει τη *δυνατότητα* και *την κατανομή της* υπό δηλωμένες παραδοχές, όχι τη συχνότητα πραγματικών επιθέσεων.

Τι δεν δείχνει: Ότι πρέπει να εγκαταλειφθεί η ιατρική ΤΝ· ότι κάθε μοντέλο διαρρέει ή ότι κάποιο συγκεκριμένο σύστημα σε χρήση έχει παραβιαστεί· ότι εκτίθεται το *περιεχόμενο* των ιατρικών φακέλων (και όχι η συμμετοχή)· έναν μοναδικό ποσοτικοποιημένο δείκτη ανισότητας — το ισχυρό αποτέλεσμα είναι το μοτίβο, όχι ένας αριθμός.

Κύριοι περιορισμοί: Μετρά κίνδυνο χειρότερης περίπτωσης με επιθέσεις που υλοποίησαν οι ίδιοι οι συγγραφείς, όχι παρατηρημένα περιστατικά· οι δραματικοί αριθμοί αφορούν εκ σχεδιασμού τα πιο εκτεθειμένα άτομα· και οι ακριβείς ανισότητες ανά ομάδα παρουσιάζονται ως σταθερό μοτίβο σε πολλά σύνολα δεδομένων αντί να συμπυκνώνονται σε έναν αριθμό.

Πόση εμπιστοσύνη πρέπει να έχει ένας γενικός αναγνώστης; Υψηλή ότι οι συγκεντρωτικές μετρικές ιδιωτικότητας υποτιμούν τον ατομικό κίνδυνο, ότι ο υπολειπόμενος κίνδυνος συγκεντρώνεται στους υποεκπροσωπούμενους ασθενείς και ότι επιδεινώνεται όσο μεγαλώνει το μοντέλο — αυτό έχει αποδειχθεί σε πολλά σύνολα δεδομένων και μοντέλα. Μέτρια όσον αφορά τη συχνότητα τέτοιων επιθέσεων στον πραγματικό κόσμο, την οποία η μελέτη δεν μετρά. Η ασφαλής ανάγνωση δεν είναι «η ιατρική ΤΝ διαρρέει τα δεδομένα σου», ούτε «η ιδιωτικότητα έχει λυθεί», αλλά: «ο τυπικός έλεγχος ιδιωτικότητας είναι τυφλός στις χειρότερες περιπτώσεις του, και αυτές αφορούν τους πιο ευάλωτους ασθενείς — άρα ο έλεγχος πρέπει να αλλάξει».

Πηγές

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