The Clean Paper

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

AI רפואי יכול להיות פרטי בממוצע ועדיין לחשוף מטופלים מסוימים — ובעיקר את אלה שאינם מיוצגים מספיק

הטענה שמודל AI רפואי הוא "שומר פרטיות" נשענת בדרך כלל על מספר ממוצע אחד. המחקר הזה טוען שזה המבחן הלא נכון. החוקרים בחנו מתקפות הסקת השתייכות — שמנסות לגלות אם הרשומה של אדם מסוים הייתה בנתוני האימון, ולכן עשויות להסגיר למשל שהיה לו מצב רפואי מסוים — ומדדו את הסיכון לכל מטופל בנפרד, בשבעה מערכי נתונים רפואיים (הדמיה, ECG ורשומות אלקטרוניות) ובמודלים רבים. התבנית ברורה. מודלים שנראים בטוחים בממוצע יכולים עדיין לאפשר זיהוי כמעט מושלם של אנשים מסוימים (AUC של המתקפה ≥ 0.95); החשופים הם באופן שיטתי מקבוצות בתת-ייצוג, והבעיה מחמירה ככל שהמודלים גדלים. המחברים אינם קוראים לנטוש את AI רפואי, אלא למדוד פרטיות ברמת המטופל, לשלוט בגישה למודל ולהשתמש בפרטיות דיפרנציאלית. הליבה הלא-נוחה: "פרטי בממוצע" אינו הבטחת פרטיות, והכשל פוגע דווקא במטופלים שכבר מוגנים פחות.

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

הבטחת פרטיות היא הבטחה לאדם, לא לממוצע

כאשר אומרים שמודל בינה מלאכותית רפואית הוא "שומר פרטיות", הטענה נשענת בדרך כלל על מספר אחד: בממוצע, על פני כל המטופלים שהנתונים שלהם שימשו לאימון המודל, הסיכוי להסיק אם אדם מסוים נכלל במערך האימון נמוך. זה נשמע מרגיע. הנקודה השקטה והלא-נוחה של המאמר הזה היא שזה המספר הלא נכון.

פרטיות איננה ממוצע. זו הבטחה לכל אדם ואדם — שהעובדה שהנתונים שלו נכללו במערך האימון לא תחזור ותחשוף אותו. ממוצע יכול לקיים את ההבטחה כמעט לכולם ובו בזמן להפר אותה לחלוטין עבור מעטים. החוקרים מדדו את סיכון הפרטיות ברמת המטופל הבודד, ברזולוציה שלא נעשה בה שימוש קודם, ומצאו בדיוק את זה: מודלים שנראים בטוחים כשמסתכלים עליהם במצטבר יכולים לדלוף כמעט באופן מושלם את עצם השתייכותם של אנשים מסוימים למערך האימון — והאנשים שנחשפים כך הם, באופן לא-פרופורציונלי, אלה שמלכתחילה מוגנים פחות.

מה עשו החוקרים

הצוות — בהובלת Knolle, Moritz עם Rückert Daniel (שניהם מהאוניברסיטה הטכנית של מינכן) ו-Kaissis Georg (ממכון Plattner), Hasso לצד עמיתים מ-London College Imperial — לקח איום פרטיות מוכר בשם **מתקפת הסקת השתייכות** (membership inference attack) ושינה את השאלה. במקום לשאול "באיזו תדירות, בממוצע, המתקפה מצליחה בכל מערך הנתונים?", הם שאלו: "עד כמה המטופל המסוים הזה חשוף?"

הם הפעילו את הניתוח ברמת המטופל על **שבעה מערכי נתונים רפואיים מבוססים**, המכסים סוגים שונים מאוד של מידע — הדמיה רפואית, אלקטרוקרדיוגרמות ורשומות רפואיות אלקטרוניות — ובכל אחד מהם בחנו 200 מודלים. עבור כל מטופל בכל מודל הם העריכו באיזו רמת ביטחון תוקף יכול לקבוע אם הרשומה של אותו אדם הייתה חלק מנתוני האימון, ולאחר מכן פילחו את התוצאות לפי קבוצות: מצב מחלה, גזע בדיווח עצמי, ביטוח, מין ופרוטוקול הדמיה.

מהי בעצם מתקפת הסקת השתייכות

העובדה עצם — השתייכותל- נוגעת היא שלך". הרשומה את פולט מ"המודל יותר עדינה כאן הדליפה האימון. במערך היו שלך שהנתונים

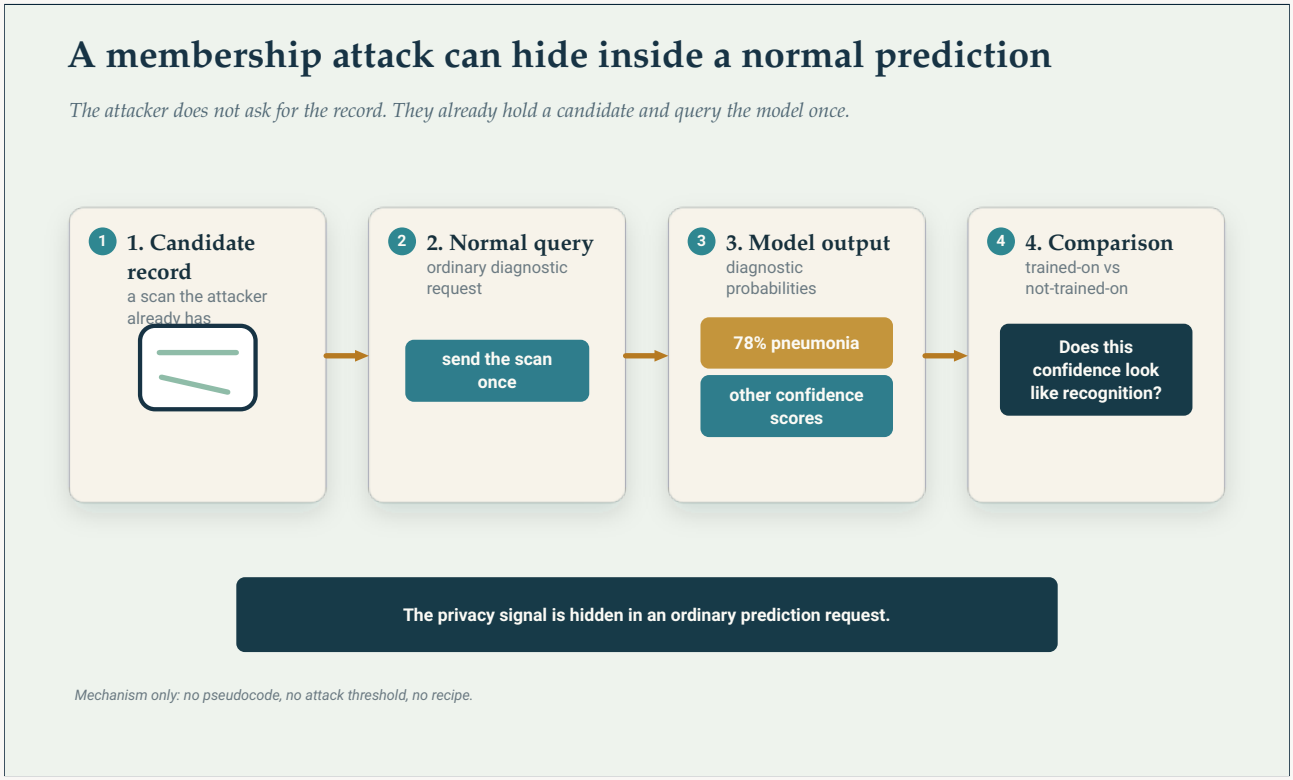
נתחיל ממה שמודל כזה עושה בדרך כלל. נותנים לו סריקה — למשל צילום חזה — והוא מחזיר הסתברויות: 78% סיכוי לדלקת ריאות, יחד עם הערכות לקרדיומגליה, בצקת, קונסולידציה וכדומה. התשובה האבחונית היומיומית הזאת היא כל מה שהמתקפה משתמשת בו. אין כאן ממשק סודי או גישה חריגה.

החולשה שעליה היא נשענת היא זו: מודל נוטה להיות מעט **בטוח יותר** לגבי הדוגמאות המדויקות שעליהן אומן מאשר לגבי דוגמאות שמעולם לא ראה. אפשר לחשוב על תלמיד שראה בסתר את המבחן מראש — בשאלות שכבר תרגל, התשובות מגיעות מעט מהר מדי ובביטחון מעט גבוה מדי. להסיק מהסימן הזה אם רשומה מסוימת הייתה אחת הדוגמאות שהמודל למד מהן — זה בדיוק מה פירוש "membership inference". כך נראית המתקפה, שלב אחר שלב. מישהו רוצה לדעת אם סריקה של אדם מסוים שימשה לאימון המודל. הוא **לא** מבקש מהמודל להחזיר את הרשומה — כבר יש בידיו סריקה מועמדת, של אותו אדם או עותק קרוב שלה. הוא שולח אותה פעם אחת, כמו בקשת אבחון רגילה, ורושם עד כמה המודל בטוח בתשובתו. לאחר מכן הוא בודק אם רמת הביטחון הזאת דומה יותר למודל שאומן על הסריקה או למודל שלא אומן עליה — השוואה שאפשר לבצע בעלות נמוכה יחסית באמצעות אימון מודל חלופי משלו ("reference model"), כדי ללמוד איך נראית אותה עודפות ביטחון אופיינית, על מחשב רגיל וללא חומרה מיוחדת. אם המודל האמיתי בטוח באופן

חריג לגבי הסריקה הזאת — כאילו הוא "מזהה" אותה — זו ראיה חזקה לכך שהסריקה הייתה בנתוני האימון. החלק המטריד הוא ששום דבר בבקשה לא נראה כמו מתקפה: זו אותה שאילתת אבחון שרופא עשוי לבצע, ואות הפרטיות חבוי בתוך תחזית רגילה. בדיוק משום כך הסיכון ממשי — אף שעד כה הוא הודגם בתנאי מעבדה ובכפוף להנחות שהוגדרו, ולא זוהה כמתקפה בעולם האמיתי.

למה בכלל משנה ההשתייכות, אם הרשומה עצמה אינה דולפת? כי עצם ההשתייכות היא עובדה. אם מודל אומן על מטופלים שקיבלו אימונותרפיה מסוימת לסרטן, אישור לכך שהרשומה שלך נכללה בו יכול לחשוף שסביר שהיה לך אותו סרטן — בדיוק סוג המידע שמבטח או מעסיק לא אמורים להיות מסוגלים להסיק. תוכן הרשומה נשאר סגור; מה שבורח הוא עצם השייכות לקבוצה.

המחברים מציגים במפורש את ההנחות — גישה לתחזיות הרגילות של המודל, רשומה מועמדת ומודל ייחוס של התוקף — לא כמדריך ביצוע, אלא כדי שאפשר יהיה לנתח את האיום באופן ענייני ולא לנפנף אותו הצידה.



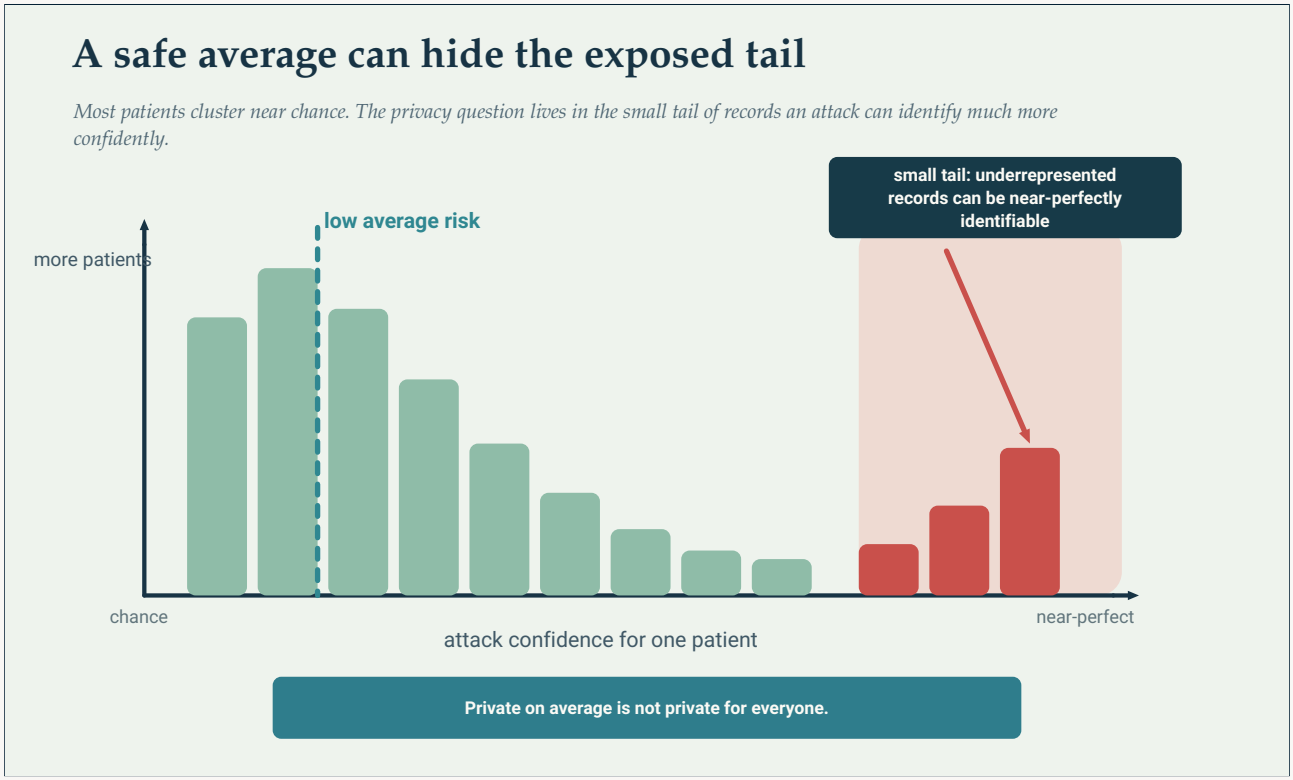
המתקפה אינה זקוקה ל-API פרטיות מיוחד. שאילתת אבחון רגילה יכולה לדלוף אות השתייכות אם המודל בטוח באופן חריג לגבי רשומה שכבר ראה באימון.

מה הם מצאו

שלושה ממצאים, כל אחד חד יותר מקודמו.

הממוצעים מסתירים את החשופים. כאשר מודדים במצטבר — כפי שנהוג — רבים מהמודלים נראים מרגיעים מבחינת פרטיות: אצל רוב המטופלים המתקפה אינה מצליחה יותר מניחוש אקראי. אבל התמונה ברמת המטופל

נראתה אחרת. כפי שאמר Knolle ב-הודעת המחקר, הערכות קודמות מדדו רק את הסיכון הממוצע על פני כל המטופלים; כאן הסיכון נבדק לראשונה ברמת המטופל הבודד, והתמונה השתנתה מאוד. עבור אנשים מסוימים המתקפה מצליחה **כמעט באופן מושלם** — המחברים מודדים זאת כ-AUC של המתקפה של **0.95 ומעלה** (0.5 שקול להטלת מטבע, 1.0 הוא סיווג מושלם) — גם כאשר הממוצע של מערך הנתונים כולו נראה לא טוב יותר ממקריות.



הממוצע יכול להישאר קרוב לרמת המקריות, בעוד זנב קטן של רשומות חשוף הרבה יותר. נקודת המאמר היא שצריך לבדוק פרטיות ברמת המטופל, ולא רק במצטבר.

החשיפה אינה שוויונית — והיא נופלת על קבוצות שאינן מיוצגות מספיק. המטופלים שהיו פגיעים ביותר למתקפה כמעט-מושלמת השתייכו באופן שיטתי לקבוצות **בתת-ייצוג בנתונים**: מיעוטים אתניים, פנוטיפים של מחלות נדירות או מאפייני הדמיה חריגים. במערך הרשומות האלקטרוניות, מטופלים שחורים הופיעו בין הרשומות הפגיעות ביותר בתדירות גבוהה ב-31% מהצפוי; במערך הממוגרפיה, סריקות שסומנו כחשודות לממאירות היו מיוצגות באזור הסיכון הזה ביתר עצום של +1,179%. (שני המספרים האלה הם דוגמאות, לא התמונה כולה — המאמר מדווח על אותו דפוס במספר קבוצות — והם מתארים ייצוג-יתר יחסי בתוך הזנב הקטן של הסיכון הקיצוני: כמה פעמים יותר מהצפוי מטופלים אלה מופיעים בין הרשומות החשופות ביותר, ולא איזה שיעור מכלל המטופלים השחורים או מכלל הממוגרפיות החשודות חשוף.) למודל יש פחות דוגמאות דומות ש"יטשטשו" מטופלים כאלה בתוך הקבוצה, ולכן הרשומות שלהם בולטות יותר — ובדיוק את החריגות הזאת מתקפת השתייכות מסוגלת לזהות. כשל הפרטיות אינו אקראי; הוא מתרכז באנשים שכבר נמצאים בשוליים.

מודלים גדולים יותר מחמירים את הבעיה. מספר המטופלים שנחשפו למתקפות כמעט-מושלמות עלה בחדות עם קיבולת המודל — במערך נתונים דרמטולוגי אחד, השיעור טיפס מכמעט אפס במודל הקטן ביותר לכ-**אחד מכל עשרה** במודל הגדול ביותר. ככל שמודלים רפואיים נעשים גדולים ומסוגלים יותר — הכיוון שבו כל התחום מתקדם — הסיכון הספציפי הזה, מזהירים המחברים, מחמיר ולא משתפר.

הסיכום של Rückert חד: זה אינו סיכון שאפשר לקבל; נתוני בריאות רגישים מאוד.

למה "בטוח בממוצע" הוא המבחן הלא נכון

מפתה לקרוא סיכון ממוצע נמוך כאישור לכך שהכול בסדר. הלקח המרכזי של המאמר הוא שממוצע כאן אינו רק מדד גס — הוא מודד את הדבר הלא נכון.

הבטחת פרטיות משמעותית רק אם היא תקפה גם לאדם שנמצא בסיכון הגבוה ביותר, לא רק לאדם שבאמצע ההתפלגות. מודל שבו 999 מתוך 1,000 מטופלים אינם ניתנים לזיהוי אך מטופל אחד ניתן לזיהוי כמעט מושלם אינו "פרטי ב-99.9%" במובן שיש לו משמעות עבור אותו אדם — ואם אותו אדם הוא באופן צפוי בעל מחלה נדירה או בן מיעוט אתני, המדד אינו רק חלקי, אלא גם מסתיר אי-שוויון. הוא מודד בטיחות עבור הרוב וקורא לה בטיחות עבור כולם.

זה השינוי שהמאמר כופה: מ-עד כמה המודל הזה פרטי בממוצע? ל-מי המטופל החשוף ביותר, ומי הוא? אלה שאלות שונות, ורק השנייה היא באמת שאלת פרטיות.

מה המחקר לא מוכיח — ולא טוען

- הוא **לא** אומר שצריך לנטוש בינה מלאכותית רפואית. המסגור של המחברים הוא הפחתת סיכון, לא נסיגה: למדוד ולתקן את הבעיה, לא להפסיק לפתח.
- הוא **לא** אומר שכל מודל רפואי דולף, או שמודל מסוים שכבר פרוס הותקף. הוא מראה שה-סיכון קיים ומתחלק באופן לא שוויוני, ושמדדים מצטברים רגילים מפספסים אותו.
- הוא **לא** אומר שהרשומות שלך כבר חשופות. המתקפה דורשת תנאים מסוימים — גישה למודל, רשומה מועמדת ותשתית של התוקף — ולא יכולת מזדמנת.
- הוא **לא** מראה שתוכן הרשומה של המטופל דולף. מה שדולף הוא ההשתייכות — עצם העובדה שהרשומה נכללה — וזה מסוכן מסיבה אחרת, לא מפני שהרשומה עצמה נשלפת.
- הוא **לא** מצמצם את אי-השוויון למספר כותרת יחיד. התוצאה החזקה והחוזרת היא ה-דפוס: מדדים מצטברים ממעיטים בסיכון האישי, והמטופלים שאינם מיוצגים דיים נושאים בחלק הגדול ביותר ממנו — לאורך מערכי נתונים ומאות מודלים.

עד כמה הראיות חזקות?

זו תוצאה אמפירית ומתודולוגית, והיא מבוססת היטב. הדפוס — מדדי פרטיות מצטברים ממעיטים באופן שיטתי בסיכון ברמת המטופל, הסיכון שנותר מתרכז בקבוצות בתת-ייצוג, והוא מחמיר עם קיבולת המודל — נשמר על פני **שבעה מערכי נתונים מסוגים שונים** ומספר גדול של מודלים בכל מערך. הרוחב הזה הוא בדיוק מה שהופך את טענת המדידה לאמינה יותר מאשר ארטיפקט של מערך נתונים יחיד.

שתי הסתייגויות חשובות. ראשית, זו הדגמה של סיכון, שנמדד באמצעות המתקפות שהחוקרים עצמם הפעילו; היא מתארת עד כמה תוקף מיומן יכול להצליח בהנחות שהוגדרו, לא באיזו תדירות מתקפות כאלה מתרחשות בעולם האמיתי. שנית, המספרים הבולטים — הצלחה כמעט מושלמת עבור מטופלים מסוימים — מתארים בכוונה את האנשים שנמצאים בקצה הפגיע ביותר; זו בדיוק מטרת הניתוח, ויש לקרוא זאת כ"הזנב כבד הרבה יותר מכפי שהממוצע מרמז", לא כ"רוב המטופלים חשופים".

העמדה המתאימה היא לא בהלה ולא ביטול: זה מחקר זהיר, רב-מערכי-נתונים, שמראה שהדרך הסטנדרטית שבה

אנו מאשרים פרטיות של AI רפואי עיוורת למקרים הגרועים ביותר שלה — ושאותם מקרים נופלים על המטופלים שיש להם הכי מעט מרווח לספוג פגיעה נוספת.

למה זה חשוב

שני דברים הופכים את המחקר הזה ליותר מהערת שוליים טכנית.

ראשית, הוא משנה את מה שצריך להיות מותר לכנות "שומר פרטיות". אם מודל משוחרר עם ציון פרטיות ממוצע, הציון יכול להיות באמת נמוך ובכל זאת להסתיר תת-קבוצה של מטופלים שניתנים לזיהוי כמעט מושלם באמצעות מתקפת השתייכות. הדרישה המעשית של המאמר קונקרטית: להעריך סיכון פרטיות **לכל מטופל** לפני פריסה, לשלוט במי שיכול לגשת למודלים פרוסים, ולהשתמש בשיטות כמו **פרטיות דיפרנציאלית** (differential privacy) — הוספה של רעש קטן ומכויל בקפידה במהלך האימון, שמקהה מתקפות השתייכות במחיר אמיתי ומנוהל של ירידה מסוימת בשימושיות המודל. פשרת הפרטיות–תועלת מפורשת; היא אינה בחינם. "בדקנו את הממוצע" לא צריך להיחשב עוד כ"בדקנו את הפרטיות".

שנית, הוא קושר יחד פרטיות והוגנות. אותן קבוצות שאינן מיוצגות מספיק בנתונים רפואיים — ולכן לעיתים כבר מקבלות שירות פחות טוב מ-AI רפואי — הן גם אלה שהפרטיות שלהן מוגנת פחות בידי אותו AI. תחום שמשקיע מאמץ באי-השוויון הראשון אינו יכול להתייחס לשני כבעיה של מישהו אחר. מדובר באותם אנשים.

הסיפור המרגיע על פרטיות ב-AI רפואי — מדרגו אותה, הממוצע נמוך, הכול בסדר — הוא בדיוק הסיפור שהמאמר מפרק. לא כדי להבריח אנשים מהטכנולוגיה, אלא כדי להזיז את הרף למקום שבו הוא צריך להיות: הבטחה שאפשר לטעון שקיימת רק אם בדקנו אותה גם עבור האדם שסביר ביותר שייפגע.

בקצרה

מתקפות הסקת השתייכות מנסות לקבוע אם הרשומה של אדם מסוים הייתה בנתוני האימון של מודל AI — ומכיוון שעצם ההשתייכות יכולה לחשוף מידע (למשל שהאדם חלה במחלה מסוימת), זו דליפת פרטיות אמיתית גם כאשר תוכן הרשומה לעולם אינו נחשף. מחקרים קודמים מדדו באיזו תדירות מתקפות כאלה מצליחות בממוצע על פני מערך נתונים. המחקר הזה מדד את הסיכון לכל מטופל בנפרד, על פני שבעה מערכי נתונים רפואיים (הדמיה, ECG, רשומות אלקטרוניות) ומודלים רבים בכל אחד מהם, ומצא שמדדים מצטברים ממעיטים מאוד בסיכון: אנשים מסוימים ניתנים לזיהוי כמעט מושלם (AUC של המתקפה ≥ 0.95) גם כאשר הממוצע הכללי נראה בטוח; הפגיעים ביותר הם באופן שיטתי אנשים מקבוצות בתת-ייצוג (מיעוטים אתניים, מחלות נדירות, מאפייני הדמיה חריגים); והבעיה מחמירה ככל שהמודלים גדלים. המחברים אינם טוענים שצריך לנטוש AI רפואי — הם קוראים למדוד פרטיות ברמת האדם, לשלוט בגישה למודלים ולהשתמש בפרטיות דיפרנציאלית. המסר: "פרטי בממוצע" אינו הבטחת פרטיות, והכשל נופל בדרך כלל על האנשים שכבר מוגנים פחות.

בדיקה מפוכחת

מה המחקר מראה: על פני שבעה מערכי נתונים רפואיים ומודלים רבים בכל אחד מהם, סיכון הסקת ההשתייכות ברמת המטופל גבוה עבור אנשים מסוימים הרבה יותר מכפי שמדדים מצטברים מרמזים — עד הצלחה כמעט מושלמת

של המתקפה AUC) ≥ 0.95) — והסיכון השיורי נופל באופן לא-פרופורציונלי על קבוצות בתת-ייצוג ומחמיר עם קיבולת המודל.

מה סביר אבל לא הוכח: שתוקפים בעולם האמיתי כבר מנצלים זאת נגד מודלים קליניים פרוסים; המחקר מדגים את היכולת ואת התפלגותה בתנאים שהוגדרו, לא את שכיחות המתקפות בפועל.

מה הוא אינו מראה: שצריך לנטוש AI רפואי; שכל מודל דולף או שמודל מסוים שנפרס כבר נפרץ; שתוכן הרשומה של המטופל, ולא עצם ההשתייכות, נחשף; או מספר יחיד שמסכם את אי-השוויון — התוצאה העמידה היא הדפוס, לא מספר בודד.

המגבלות העיקריות: המחקר מודד סיכון בקצה באמצעות המתקפות של המחברים עצמם, לא תקריות שנצפו בעולם האמיתי; המספרים הדרמטיים מתארים בכוונה את האנשים החשופים ביותר; והפערים המדויקים בין קבוצות מוצגים כדפוס עקבי על פני מערכי נתונים, ולא מצומצמים למדד יחיד.

כמה ביטחון צריך להיות לקורא כללי? גבוה בכך שמדדי פרטיות מצטברים ממעיטים בסיכון אישי, שהסיכון הנותר מתרכז במטופלים שאינם מיוצגים מספיק, ושהוא מחמיר עם גודל המודל — הדבר הודגם על פני מערכי נתונים ומודלים רבים. בינוני לגבי שכיחות מתקפות כאלה בעולם האמיתי, שאותה המחקר אינו מודד. הקריאה הבטוחה היא לא "AI רפואי מדליף את הנתונים שלך", וגם לא "בעיית הפרטיות נפתרה", אלא: "בדיקת הפרטיות הסטנדרטית עיוורת למקרים הגרועים ביותר שלה, והמקרים האלה נופלים על המטופלים הפגיעים ביותר — ולכן הבדיקה חייבת להשתנות."

מקורות

(2026) Nature AI, medical from risks privacy Disparate al., et Knolle
https://doi.org/10.1038/s41586-026-10688-0

(2026) AI' medical in risks privacy reveals 'Study release, press — Munich of University Technical
https://www.tum.de/en/news-and-events/all-news/press-releases/details/study-reveals-priva

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