The Clean Paper

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

Tibbi AI orta hesabla məxfi olub yenə də konkret pasiyentləri — xüsusən az təmsil olunanları — ifşa edə bilər

"Məxfiliyi qoruyan" adlandırılan tibbi AI modeli adətən bir orta göstəriciyə söykənir. Bu tədqiqat həmin rəqəmin yanlış test olduğunu göstərir. Müəlliflər konkret şəxsin qeydinin modelin təlim məlumatlarında olub-olmadığını üzə çıxaran membership inference hücumlarını araşdıraraq riski ümumi deyil, pasiyent səviyyəsində ölçdülər: yeddi tibbi məlumat dəsti (görüntüləmə, ECG, elektron qeydlər) və hər birində çoxlu model üzrə. Nümunə belədir: orta hesabla təhlükəsiz görünən modellər bəzi konkret insanları demək olar qüsursuz müəyyən etməyə imkan verə bilər (hücum AUC-si ≥ 0.95); ifşa olunanlar sistematik olaraq az təmsil olunan qruplardandır (etnik azlıq, nadir xəstəlik, qeyri-adi görüntüləmə); model böyüdükcə risk artır. Müəlliflər tibbi AI-dan imtina etməyi yox, məxfiliyi hər pasiyent üçün ölçməyi, modelə çıxışı nəzarətdə saxlamağı və differential privacy tətbiq etməyi təklif edirlər. Narahatedici əsas fikir: "orta hesabla məxfi" məxfilik zəmanəti deyil və bu zəmanət ən az qorunan pasiyentlər üçün daha çox pozulur.

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

Məxfilik zəmanəti ortalamaya deyil, insana verilən vəddir

Tibbi süni intellekt modeli “məxfiliyi qoruyan” adlandırılanda bu iddia adətən bir rəqəmə söykənir: modeli öyrətmək üçün məlumatlarından istifadə edilmiş bütün pasiyentlər üzrə orta hesabla, hər hansı konkret şəxsin təlim dəstinə daxil olub-olmadığını müəyyənləşdirmək ehtimalı aşağıdır. Bu, sakitləşdirici səslənir. Bu məqalənin sakit, amma narahatedici fikri budur ki, yanlış rəqəmə baxırıq.

Məxfilik orta göstərici deyil. O, hər bir insana ayrıca verilən vəddir — təlim dəstində olmaq sonradan onun əleyhinə işləyib onu ifşa etməyəcək. Orta göstərici bu vədi demək olar hamı üçün saxlaya, amma bir neçə nəfər üçün tam poza bilər. Tədqiqatçılar məxfilik riskini əvvəllər istifadə olunmamış dəqiqliklə pasiyent-pasiyent ölçdülər və məhz bunu gördülər: ümumi göstəricilərə görə təhlükəsiz görünən modellər bəzi konkret şəxslərin təlim dəstində iştirakını demək olar qüsursuz şəkildə sızdıra bilər — və bu sızmaya məruz qalanlar qeyri-mütənasib olaraq onsuz da ən az qorunan insanlardır.

Müəlliflər nə etdilər

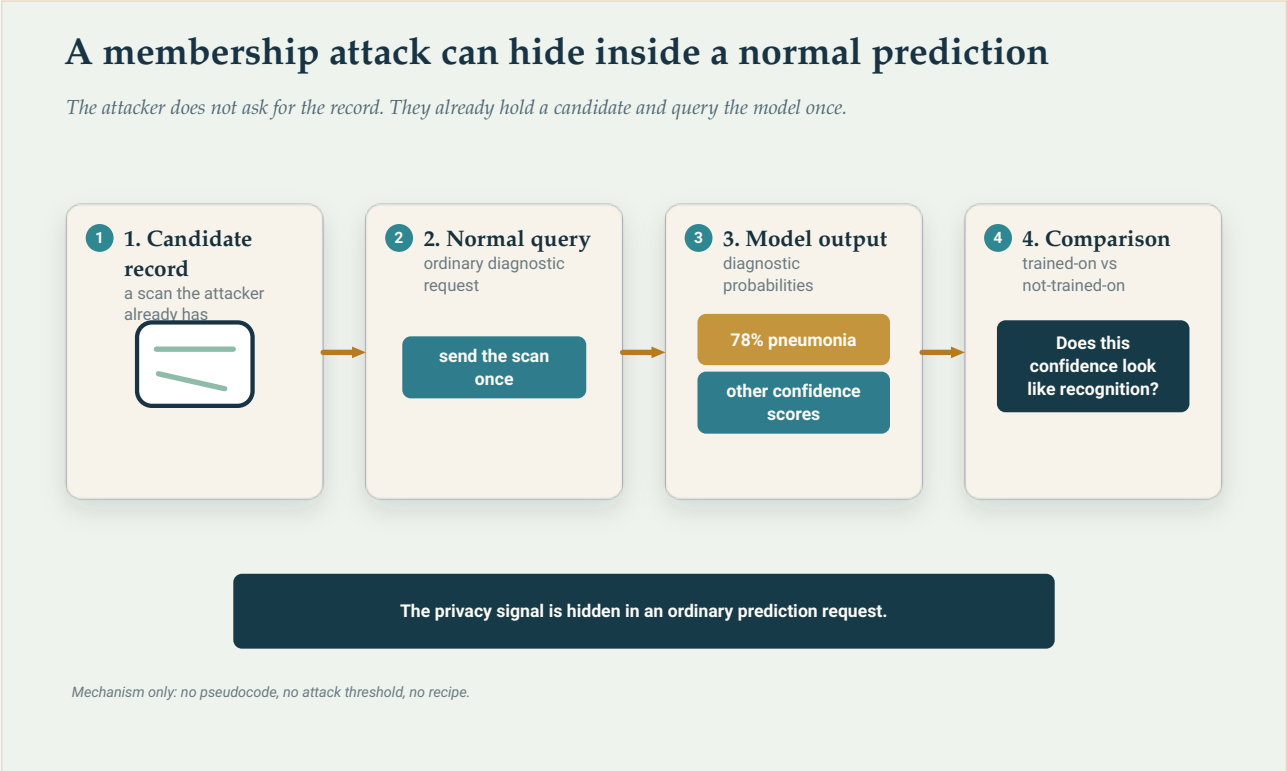
Moritz Knolle-nin rəhbərlik etdiyi komanda — Daniel Rückert (hər ikisi Münhen Texniki Universitetindən) və Georg Kaissis (Hasso Plattner Institute), həmçinin Imperial College Londondan həmkarları ilə birlikdə — **membership inference attack** adlanan məşhur məxfilik hücumunu götürüb sualı dəyişdirdi. “Bu hücum bütün məlumat dəsti üzrə orta hesabla nə qədər tez-tez uğur qazanır?” əvəzinə “*bu konkret pasiyent* nə dərəcədə açıq qalır?” deyə soruşdular.

Onlar bu pasiyent-səviyyəli təhlili çox fərqli məlumat növlərini — tibbi görüntüləmə, elektrokardioqramlar və elektron sağlamlıq qeydlərini — əhatə edən **yeddi tanınmış tibbi məlumat dəsti** üzrə və hər birində 200 model üzərində apardılar. Hər modeldə hər pasiyent üçün hücumçunun həmin şəxsin qeydinin təlim məlumatlarının bir hissəsi olub-olmadığını nə qədər əminliklə müəyyən edə biləcəyini hesabladılar, sonra nəticələri qruplara görə ayırdılar: xəstəlik statusu, şəxsin özü tərəfindən bildirilən irq, sığorta, cins və görüntüləmə protokolu.

Membership inference hücumu əslində nədir

Buradakı sızma “model sizin tibbi qeydinizi olduğu kimi çıxarır” fikrindən daha incədir. Söhbət *üzvlükdən* — məlumatınızın təlim dəstində olmasının özündən gedir.
Əvvəlcə belə modellərdən birinin normalda nə etdiyini düşünək. Ona, məsələn, döş qəfəsinin rentgen şəklini verirsiniz və o sizə ehtimallar qaytarır: pnevmoniya ehtimalı 78%, üstəlik kardiomeqaliya, ödem və konsolidasiya üçün göstəricilər. Hücumun istifadə etdiyi *yalnız* bu adi diaqnostik cavabdır. Heç bir ekzotik mexanizm yoxdur.
İstifadə olunan zəiflik budur: model adətən təlim zamanı gördüyü konkret nümunələrə əvvəllər heç görmədiyi nümunələrdən bir qədər **daha əmin** cavab verir. İmtahan

> suallarını əvvəlcədən gizlicə görmüş tələbəni təsəvvür edin — məşq etdiyi suallarda cavablar bir az həddən artıq tez, bir az həddən artıq əmin gəlir. Bu işarədən istifadə edib konkret bir qeydin modelin öyrəndiyi nümunələr arasında olub-olmadığını müəyyənləşdirmək "membership inference" deməkdir.
>
> Hücum addım-addım belə işləyir. Kimsə konkret şəxsin skanının modeli öyrətmək üçün istifadə olunub-olunmadığını bilmək istəyir. Onlar modeldən qeydin özünü **istəmirlər** — namizəd skan artıq əllərindədir (şəxsin öz skanı və ya ona çox yaxın surət). Onu bir dəfə adi diaqnostik sorğu kimi modelə göndərir və cavabın nə qədər əmin olduğuna baxırlar. Sonra bu əminliyin skanı *görərək təlim keçmiş* modelə, yoxsa *görməmiş* modelə daha çox bənzədiyini yoxlayırlar. Bu müqayisəni xüsusi avadanlıq olmadan adi kompüterdə öz əvəzedici modellərini ("reference model") öyrədib həmin xarakterik həddən artıq əminliyin necə göründüyünü müəyyən etməklə ucuz başa gətirə bilərlər. Əsl model bu skan barədə qeyri-adi dərəcədə əmindirsə — sanki onu tanıyırsa — bu, skanın təlim məlumatlarında olduğuna güclü dəlildir.
>
> Narahatedici tərəf odur ki, sorğunun özündə hücuma bənzəyən heç nə yoxdur: klinisistin verəcəyi adi diaqnostik sorğudur, məxfilik siqnalı isə adi proqnozun içində gizlənir. Riskin konkret olmasının səbəbi də məhz budur — baxmayaraq ki, indiyə qədər o, real mühitdə tutulmuş hadisə kimi yox, açıqlanmış laboratoriya fərziyyələri altında nümayiş etdirilib.
>
> Qeydin özü sızmırsa, üzvlük niyə vacibdir? Çünki *üzvlüyün özü faktdır*. Model, məsələn, müəyyən xərçəng immunoterapiyası alan pasiyentlərin məlumatları ilə öyrədilibsə, sizin qeydinizin orada olduğunu təsdiqləmək böyük ehtimalla həmin xərçəngə sahib olduğunuzu üzə çıxarır — sığorta şirkətinin və ya işəgötürənin heç vaxt çıxara bilməməli olduğu nəticə. Məzmun bağlı qalır; çölə çıxan, həmin qrupa aid olma faktıdır. Müəlliflər bu fərziyyələri açıq göstərirlər — modelin adi proqnozlarına çıxış, namizəd qeyd və hücumçunun öz reference modeli — bunu addım-addım hücum təlimatı kimi yox, təhlükəni əl yelləyib keçmək əvəzinə məntiqli şəkildə qiymətləndirmək üçün edirlər.



Hücum üçün xüsusi məxfilik API-si lazım deyil. Model əvvəllər gördüyü bir qeyd üçün qeyri-adi dərəcədə əmin cavab verirsə, adi diaqnostik sorğunun özü üzvlük siqnalını sızdıra bilər.

Nə tapdılar

Bir-birindən daha kəskin üç nəticə.

Ortalamalar ən çox açıq qalanları gizlədir. Ümumi şəkildə — adətən edildiyi kimi — ölçəndə bu modellərin çoxu kifayət qədər məxfi görünür: pasiyentlərin əksəriyyətində hücum təsadüfi təxmindən yaxşı deyil. Pasiyent-səviyyəli baxış isə başqa mənzərə göstərdi. Knolle tədqiqatın elanında dediyi kimi, əvvəlki qiymətləndirmələr “yalnız bütün pasiyentlər üzrə orta riski ölçüb. Biz ilk dəfə riski ayrı-ayrı pasiyentlər səviyyəsində araşdırdıq — və bu, çox fərqli mənzərə yaradır.” Bəzi şəxslər üçün hücum **demək olar qüsursuz** işləyir — müəlliflər bunu hücum AUC-si **0.95 və ya daha yüksək** kimi ölçürlər (0.5 sikkə atmaq kimidir, 1.0 qüsursuz nəticədir) — halbuki bütün məlumat dəsti üzrə orta göstərici təsadüfdən yaxşı görünmür.

Orta göstərici təsadüf səviyyəsinə yaxın qala bilər, amma qeydlərin kiçik bir "quyruğu" xeyli daha çox ifşa oluna bilər. Məqalənin əsas fikri budur ki, məxfilik yalnız ümumi deyil, pasiyent səviyyəsində də yoxlanmalıdır.

İfşa bərabər deyil — və az təmsil olunanların üzərinə düşür. Demək olar qüsursuz hücuma ən həssas pasiyentlər sistematik şəkildə məlumatlarda **az təmsil olunan qruplara** aid idi: etnik azlıqlar, nadir xəstəlik fenotipləri, qeyri-adi görüntüləmə xüsusiyyətləri. Elektron qeyd məlumat dəstində qaradərili pasiyentlər ən həssas qeydlər arasında gözləniləndən **31% daha çox** görünürdü; mammoqrafiya dəstində bədxassəlilik baxımından şübhəli işarələnmiş skanlar həmin təhlükəli zonada heyrətamiz **+1,179%** artıq təmsil olunurdu. (Bu iki rəqəm bütün mənzərə deyil, nümunələrdir — məqalə bir neçə qrupda eyni meyli göstərir — və bunlar kiçik, ekstremal risk "quyruğu" daxilində *nisbi* artıq təmsil olunmadır: bu pasiyentlərin ən çox ifşa olunan qeydlər arasında gözləniləndən nə qədər tez-tez göründüyünü göstərir, qaradərili və ya şübhəli mammoqrafiyalı pasiyentlərin neçə faizinin ifşa olunduğunu yox.) Modelin bu pasiyentləri çoxlu oxşar nümunələrin içində "əridəcək" materialı azdır, buna görə onların qeydləri seçilir — membership hücumunun aşkar etdiyi də məhz seçilməkdir. Məxfilik uğursuzluğu təsadüfi deyil; onsuz da kənarda qalan insanların üzərində cəmlənir.

Böyük modellər vəziyyəti pisləşdirir. Demək olar qüsursuz hücumlara açıq pasiyentlərin sayı model tutumu artdıqca kəskin yüksəldi — bir dermatologiya məlumat dəstində bu pay ən kiçik modeldə demək olar sıfırdan ən böyük modeldə təxminən **hər on nəfərdən birinə** qalxdı. Tibbi AI modelləri böyüdükcə və daha qabiliyyətli olduqca — bütün sahənin getdiyi istiqamət budur — müəlliflər xəbərdarlıq edir ki, bu konkret risk azalmaq əvəzinə artır.

Rückert-in yekunu sərtdir: "Bu, qəbul edilə bilən risk deyil. Sağlamlıq məlumatları son dərəcə həssasdır."

Niyə “orta hesabla təhlükəsizdir” yanlış testdir

Aşağı orta risk göstəricisini tam təhlükəsizlik arayışı kimi oxumaq cazibədardır. Məqalənin əsas dərsi budur ki, burada ortalamaq sadəcə qeyri-dəqiq deyil — yanlış şeyi ölçür.

Məxfilik zəmanəti yalnız ortadakı insan üçün deyil, ən çox risk altında olan insan üçün də işləyirsə mənalıdır. 1,000 pasiyentdən 999-nun üzvlüyü müəyyən edilə bilməyən, amma bir nəfərin demək olar qüsursuz müəyyən edildiyi model həmin bir insan üçün vacib olan heç bir mənada “99.9% məxfi” deyil — üstəlik həmin bir nəfər proqnozlaşdırıla bilən şəkildə nadir xəstəlikli pasiyent və ya etnik azlıq nümayəndəsidirsə, metrik yalnız natamam deyil, səssizcə ayrı-seçkilik yaradır. Əksəriyyət üçün təhlükəsizliyi ölçüb bunu hamı üçün təhlükəsizlik adlandırır.

Məqalənin məcbur etdiyi dəyişiklik budur: *bu model orta hesabla nə qədər məxfidir?* sualından *ən çox ifşa olunan pasiyent kimdir və hansı qrupa aiddir?* sualına keçmək. Bunlar fərqli suallardır və yalnız ikincisi həqiqi məxfilik sualıdır.

Bu nəyi *sübut etmir* — və iddia etmir

- Tibbi AI-dan **imtina etməyi** demir. Müəlliflərin yanaşması geri çəkilmək yox, riski azaltmaqdır: riski ölç və düzəlt, qurmağı dayandırma.
- Hər tibbi modelin məlumat sızdırdığını və ya istifadədə olan hər hansı konkret modelə hücum edildiyini **demir**. Riskin *mövcud və qeyri-bərabər paylandığını*, standart ümumi metriklərin isə onu görmədiyini göstərir.
- Sizin qeydlərinizin artıq ifşa olunduğu **demək deyil**. Hücum üçün konkret şərtlər lazımdır — modelə çıxış, namizəd qeyd və hücumçunun öz infrastrukturu — bu, təsadüfi istifadəçinin adi qabiliyyəti deyil.
- Pasiyent qeydinin *məzmununun* sızdığını **göstərmir**. Sızan üzvlükdür — daxil edilmə faktı — və bu, qeydin birbaşa çıxarılmasına görə yox, başqa səbəbdən təhlükəlidir.
- Bərabərsizlikləri bircə sensasiyalı rəqəmə **endirmir**. Güclü və təkrarlanan nəticə *nümunədir*: ümumi metriklər fərdi riski az göstərir və yükün ən böyüyü az təmsil olunan pasiyentlərin üzərinə düşür — bu, məlumat dəstləri və yüzlərlə model boyunca görünür.

Dəlil nə qədər güclüdür?

Bu, empirik və metodoloji nəticədir — və kifayət qədər möhkəmdir. Ümumi məxfilik metriklərinin pasiyent-səviyyəli riski sistematik şəkildə az göstərməsi, qalan riskin az təmsil olunan qruplarda cəmlənməsi və model tutumu artdıqca pisləşməsi **müxtəlif tipli yeddi məlumat dəsti** və hər məlumat dəstində çoxlu model üzrə təkrarlandı. Ölçmə iddiasını tək bir məlumat dəstinin artefaktından fərqləndirib inandırıcı edən məhz bu əhatədir.

İki dürüst məhdudiyyət var. Birincisi, bu, müəlliflərin özlərinin qurduğu hücumları işlədərək ölçülmüş *risk* nümayişidir; açıqlanmış fərziyyələr altında bacarıqlı hücumçunun *nə edə biləcəyini* xarakterizə edir, real dünyada belə hücumların nə qədər tez-tez baş verdiyini yox. İkincisi, bəzi pasiyentlər üçün demək olar qüsursuz hücum uğuru kimi parlaq rəqəmlər *qəsdən* ən pis vəziyyətdə olan şəxsləri təsvir edir; bütün məqsəd də budur və onları "quyruq ortalamanın düşündürdüyündən xeyli ağırdır" kimi oxumaq lazımdır, "pasiyentlərin çoxu ifşa olunub" kimi yox.

Uyğun mövqe nə təşviş, nə də etinasızlıqdır: diqqətli, çoxsaylı məlumat dəstinə əsaslanan bir tədqiqat göstərir ki, tibbi AI məxfiliyini sertifikatlaşdırmaq üçün standart üsul özünün ən pis hallarını görmür — və həmin ən pis hallar ən az təhlükəsizlik payı olan pasiyentlərin üzərinə düşür.

Niyə vacibdir

Bunu texniki dipnotdan daha vacib edən iki şey var.

Birincisi, "məxfiliyi qoruyan" ifadəsinin nə deməsinə icazə verilməli olduğunu dəyişir. Model orta məxfilik göstəricisi ilə yayımlana bilər, həmin göstərici həqiqətən aşağı ola bilər və yenə də membership hücumu ilə demək olar qüsursuz müəyyən edilə bilən pasiyent altqrupunu gizlədə bilər. Məqalənin praktik tələbi konkretdir: buraxılışdan əvvəl məxfilik riskini **hər pasiyent üçün** qiymətləndirin, istifadədə olan modellərə kimin çıxışı olduğunu məhdudlaşdırın və **differential privacy** kimi üsullardan istifadə edin — təlim zamanı əlavə edilən kiçik, diqqətlə kalibrlənmiş səs-küy membership hücumlarını zəiflədir, amma modelin faydalılığına real və idarə olunan qiymət bahasına (məxfilik–faydalılıq kompromisi açıqdır, pulsuz deyil). "Ortalamanı yoxladıq" artıq "məxfiliyi yoxladıq" sayılmamalıdır.

İkincisi, məxfiliyi ədalətlə bir-birinə bağlayır. Tibbi məlumatlarda az təmsil olunan — buna görə də artıq tibbi AI tərəfindən daha pis xidmət alan — eyni qrupların məxfiliyinin də bu AI tərəfindən ən zəif qorunduğu üzə çıxır. Birinci bərabərsizliyi aradan qaldırmağa çalışan sahə ikincisini başqasının işi kimi görə bilməz. Söhbət eyni insanlardan gedir.

Tibbi AI məxfiliyi barədə rahatladıcı hekayə — *ölçdük, orta göstərici aşağıdır, deməli hər şey qaydasındadır* — bu məqalənin sökdüyü hekayədir. Texnologiyadan insanları qorxutmaq üçün yox, standartı aid olduğu yerə aparmaq üçün: ən çox zərər görə biləcək insan üçün də yoxladığınız halda "yerinə yetiririk" deyə biləcəyiniz vəd.

Qısa xülasə

Membership inference hücumları konkret şəxsin qeydinin AI modelinin təlim məlumatlarında olub-olmadığını müəyyən etməyə çalışır — və üzvlüyün özü məlumat verə bildiyi üçün (məsələn, həmin şəxsin müəyyən xəstəliyə sahib olduğunu üzə çıxarmaqla), əsas qeyd

heç vaxt görünməsə də bu, real məxfilik sızmasıdır. Əvvəlki işlər belə hücumların məlumat dəsti üzrə *orta hesabla* nə qədər tez-tez uğur qazandığını ölçürdü. Bu tədqiqat isə riski *hər pasiyent üçün*, yeddi tibbi məlumat dəsti (görüntüləmə, ECG, elektron qeydlər) və hər birində çoxlu model üzrə ölçdü və ümumi metriklərin riski ciddi şəkildə az göstərdiyini tapdı: bütün məlumat dəsti üzrə orta göstərici təhlükəsiz görünsə də bəzi şəxslər demək olar qüsursuz müəyyən edilə bilər (hücum AUC-si ≥ 0.95); ən həssaslar sistematik olaraq az təmsil olunan qruplardandır (etnik azlıqlar, nadir xəstəlik, qeyri-adi görüntüləmə); model böyüdükcə problem də artır. Müəlliflər tibbi AI-dan imtina etməyi yox, məxfiliyi fərdi səviyyədə ölçməyi, modelə çıxışı nəzarətdə saxlamağı və differential privacy istifadə etməyi təklif edirlər. Nəticə: "orta hesabla məxfi" məxfilik zəmanəti deyil və bu zəmanətin pozulduğu insanlar adətən onsuz da ən az qorunanlardır.

No-BS yoxlaması

Məqalənin göstərdiyi: Yeddi tibbi məlumat dəsti və hər birində çoxlu model üzrə, ayrı-ayrı pasiyentlər üçün membership-inference riski ümumi metriklərin düşündürdüyündən bəzi şəxslərdə xeyli yüksəkdir — hücum uğuru demək olar qüsursuza qədər çata bilər (AUC $\geq$ 0.95) — və qalan risk qeyri-mütənasib şəkildə az təmsil olunan qrupların üzərinə düşür, model tutumu artdıqca da böyüyür.

Mümkün, amma sübut olunmayan: Real hücumçuların artıq bunu istifadədə olan klinik modellərə qarşı tətbiq etməsi; tədqiqat açıqlanmış fərziyyələr altında *imkanı və onun paylanmasını* göstərir, real hücumların tezliyini yox.

Göstərmədiyi: Tibbi AI-dan imtina edilməli olduğu; hər modelin məlumat sızdırdığı və ya istifadədə olan konkret modelin pozulduğu; pasiyent qeydinin *məzmununun* (üzvlük yox) ifşa olduğu; bərabərsizliyi ifadə edən vahid rəqəm — möhkəm nəticə bir rəqəm yox, nümunədir.

Əsas məhdudiyyətlər: Real insidentləri yox, müəlliflərin öz hücumları ilə ən pis hal riskini ölçür; dramatik rəqəmlər qəsdən ən çox ifşa olunan şəxsləri təsvir edir; qruplar üzrə dəqiq fərqlər vahid rəqəmə endirilmək əvəzinə məlumat dəstləri boyunca ardıcıl nümunə kimi göstərilir.

Ümumi oxucu nə qədər əmin ola bilər? Ümumi məxfilik metriklərinin fərdi riski az göstərdiyinə, qalan riskin az təmsil olunan pasiyentlərdə cəmləndiyinə və model böyüdükcə pisləşdiyinə yüksək dərəcədə əmin olmaq olar — bu, çoxlu məlumat dəsti və model üzərində göstərilib. Belə hücumların real dünyada nə qədər tez-tez baş verdiyinə dair əminlik isə ortadır, çünki tədqiqat bunu ölçmür. Təhlükəsiz oxunuş belədir: nə "tibbi AI məlumatlarınızı sızdırır", nə də "məxfilik həll olunub"; "standart məxfilik yoxlaması özünün ən pis hallarını görmür və həmin hallar ən həssas pasiyentlərin üzərinə düşür — deməli, yoxlama üsulu dəyişməlidir."

Mənbələr

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