The Clean Paper

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

Bir tıbbi yapay zekâ ortalamada mahrem görünüp belirli hastaları yine de açıkta bırakabilir — özellikle az temsil edilenleri

'Mahremiyeti koruyan' denilen bir tıbbi yapay zekâ modeli genellikle tek bir ortalama sayıya dayanır. Bu çalışma o sayının yanlış test olduğunu savunuyor. Belirli bir kişinin kaydının modelin eğitim verisinde olup olmadığını açığa çıkaran ve dolayısıyla kişinin belirli bir hastalığa sahip olduğunu ele verebilen üyelik çıkarım saldırılarını inceleyen yazarlar, yedi tıbbi veri setinde (görüntüleme, EKG, elektronik kayıtlar) ve her birinde birçok modelde riski toplu değil hasta başına ölçtü. Örüntü şu: ortalamada güvenli görünen modeller, belirli kişileri neredeyse kusursuz biçimde tanımlamaya izin verebilir (saldırı AUC ≥ 0,95); açıkta kalanlar sistematik olarak az temsil edilen gruplardan gelir (etnik azınlık, nadir hastalık, alışılmadık görüntüleme); model büyüdükçe sorun ağırlaşır. Yazarlar tıbbi yapay zekâyı bırakın demiyor — mahremiyeti hasta başına ölçün, model erişimini kontrol edin ve differential privacy kullanın diyor. Rahatsız edici öz: 'ortalamada mahrem' bir mahremiyet garantisi değildir ve zaten en az korunan hastalarda başarısız olur.

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

Mahremiyet garantisi ortalamaya değil, kişiye verilmiş bir sözdür

Bir tıbbi yapay zekâ modeline "mahremiyeti koruyor" denildiğinde bu iddia çoğu zaman tek bir sayıya dayanır: modeli eğiten tüm hastalar arasında, herhangi bir kişinin eğitim kümesinde yer aldığının çıkarılabilme olasılığı ortalama olarak düşüktür. Bu güven verici gelebilir. Bu makalenin sessiz ve rahatsız edici noktası şu: yanlış sayıya bakıyoruz.

Mahremiyet ortalama değildir. Her bireye verilmiş bir sözdür — eğitim kümesinde yer almanın daha sonra dönüp onu ifşa etmeyeceği sözü. Bir ortalama bu sözü neredeyse herkese tutarken birkaç kişi için tamamen bozabilir. Araştırmacılar mahremiyet riskini daha önce kullanılmamış bir çözünürlükte, hasta hasta ölçtü ve tam olarak bunu buldu: toplu ölçütte güvenli görünen modeller belirli kişilerin üyeliğini neredeyse kusursuz biçimde sızdırabiliyor — ve sızdırdıkları kişiler orantısız biçimde zaten en az korunanlar.

Yazarlar ne yaptı?

Moritz Knolle liderliğinde, Daniel Rückert (ikisi de Technical University of Munich) ve Georg Kaissis'in (Hasso Plattner Institute), Imperial College London'dan meslektaşlarıyla birlikte oluşturduğu ekip, **membership inference attack** (üyelik çıkarım saldırısı) adı verilen iyi bilinen bir mahremiyet tehdidini alıp soruyu değiştirdi. "Bu saldırı veri setinin tamamında ortalama olarak ne sıklıkla başarılı oluyor?" yerine "*Bu belirli hasta* ne kadar açıkta?" diye sordular.

Hasta başına analizi, tıbbi görüntüleme, elektrokardiyogram ve elektronik sağlık kayıtları gibi çok farklı veri türlerini kapsayan **yedi yerleşik tıbbi veri setinde** ve her birinde 200 model üzerinde yürüttüler. Her modeldeki her hasta için bir saldırganın o kişinin kaydının eğitim verisinin parçası olup olmadığını ne kadar güvenle anlayabileceğini tahmin ettiler; ardından sonuçları hastalık durumu, kişinin bildirdiği ırk, sigorta, cinsiyet ve görüntüleme protokolü gibi gruplara göre ayırdılar.

> **Üyelik çıkarım saldırısı aslında nedir?**
>
> Buradaki sızıntı "model kaydınızı aynen dışarı döküyor"dan daha ince. Konu *üyelik* — verinizin eğitim kümesinde bulunmuş olması gerçeği.
> Önce böyle bir modelin normalde ne yaptığına bakalım. Modele bir tarama — örneğin akciğer grafisi — verirsiniz ve o olasılıklar döndürür: %78 zatürre olasılığı; yanında kardiyomegali, ödem, konsolidasyon için skorlar. Saldırının kullandığı *tek* şey bu günlük tanısal yanıttır. Egzotik bir erişim gerekmez.
> Yararlanılan zayıflık şudur: bir model genellikle üzerinde eğitim aldığı **tam örnekler** hakkında daha önce hiç görmediği örneklere göre biraz daha **emin** olur. Sınavı gizlice önceden görmüş bir öğrenci düşünün — çalıştığı sorularda yanıtlar biraz fazla hızlı, biraz fazla kendinden emin gelir. Bu ipucundan yola çıkarak belirli bir kaydın modelin çalıştığı örneklerden biri olup olmadığını bulmaya "üyelik çıkarımı" denir.

Saldırı adım adım şöyle işler. Birisi belirli bir kişinin taramasının modeli eğitmek için kullanılıp kullanılmadığını bilmek istiyor. Modelden kaydı **istemez** — aday tarama zaten elindedir (kişinin kendi taraması veya çok yakın bir kopya). Bunu sıradan bir tanı isteği gibi modele bir kez gönderir ve yanıtın ne kadar emin olduğuna bakar. Sonra bu güven düzeyinin, taramayla eğitim *yapmış* bir modele mi yoksa *yapmamış* bir modele mi daha çok benzediğini kontrol eder — bunu özel donanım gerektirmeyen sıradan bir bilgisayarda kendi vekilini ("referans model") eğiterek, bu fazla-güven işaretinin nasıl göründüğünü öğrenip ucuza yapabilir. Gerçek model bu tarama hakkında alışılmadık derecede eminse — sanki onu tanıyormuş gibi — taramanın eğitim verisinde olduğuna dair güçlü kanıt oluşur.

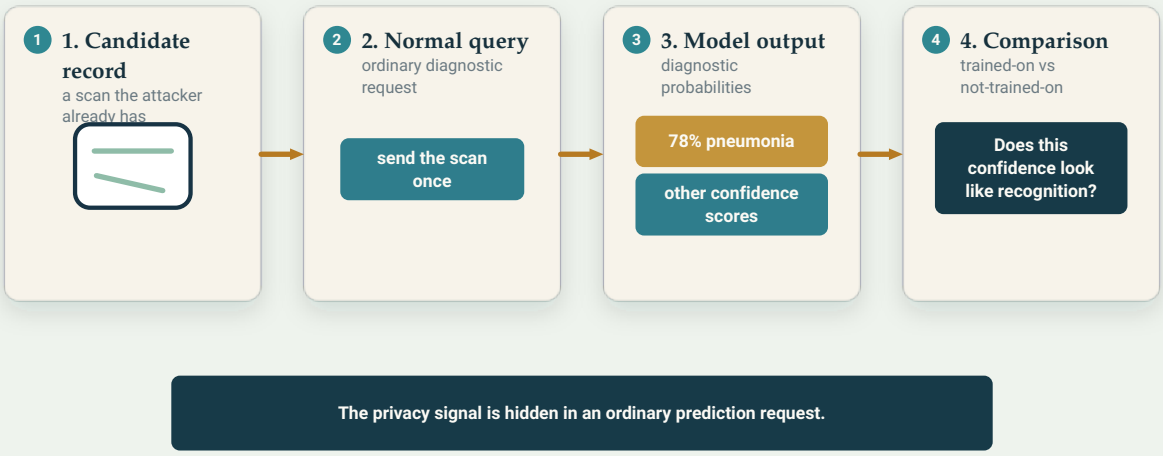
Rahatsız edici taraf, isteğin hiçbir yönünün saldırıya benzememesidir: bir klinisyenin yapacağı aynı tanı sorgusudur ve mahremiyet sinyali sıradan tahminin içine gizlenmiştir. Riskin somut olmasının nedeni tam da budur — şimdiye kadar gerçek dünyada yakalanmış saldırılar olarak değil, belirtilmiş laboratuvar varsayımları altında gösterilmiş olsa bile.

Kaydın kendisi hiç sızmıyorsa üyelik neden önemli? Çünkü *üyelik de bir bilgidir*. Bir model belirli bir kanser immünoterapisi alan hastalar üzerinde eğitildiyse, kaydınızın modelde bulunduğunun doğrulanması muhtemelen o kansere sahip olduğunuzu açığa çıkarır — bir sigortacının veya işverenin asla çıkaramaması gereken türden bilgi. İçerik kapalı kalır; dışarı kaçan ait olma gerçeğidir.

Yazarlar bu varsayımları — modelin sıradan tahminlerine erişim, aday bir kayıt ve saldırganın kendi referans modeli — bir kullanım kılavuzu vermek için değil, tehdidin el sallayarak geçiştirilmek yerine akıl yürütülebilir olması için açıkça belirtiyor.

A membership attack can hide inside a normal prediction

The attacker does not ask for the record. They already hold a candidate and query the model once.



Mechanism only: no pseudocode, no attack threshold, no recipe.

Saldırı özel bir mahremiyet API'sine ihtiyaç duymaz. Model daha önce gördüğü bir kayıt üzerinde olağandışı derecede eminse sıradan bir tanı sorgusu bile üyelik sinyali sızdırabilir.

Ne buldular?

Üç bulgu var ve her biri öncekinden daha keskin.

Ortalamalar açıkta kalanları gizliyor. Toplu olarak — alışılmış biçimde — ölçüldüğünde bu modellerin çoğu rahatlatıcı derecede mahrem görünüyor: saldırı hastaların çoğunda yazı-turadan daha iyi değil. Hasta başına görünüm farklı bir hikâye anlattı. Knolle'nin çalışmanın duyurusunda söylediği gibi, önceki değerlendirmeler "yalnızca tüm hastalar arasındaki ortalama riski ölçmüştü. Biz ilk kez tek tek hasta düzeyindeki riski inceledik — ve çok farklı bir tablo ortaya çıkıyor." Bazı bireylerde saldırı **neredeyse kusursuz** çalışıyor — yazarlar bunu saldırı AUC'sinin **0,95 veya üzeri** olmasıyla ölçüyor (0,5 yazı-tura, 1,0 kusursuz) — veri seti genelindeki ortalama ise şanstan daha iyi görünmese bile.

Ortalama şans düzeyine yakın kalırken küçük bir kayıt kuyruğu çok daha fazla açıkta olabilir. Makalenin noktası, mahremiyetin yalnızca toplu olarak değil hasta düzeyinde kontrol edilmesi gerektiğidir.

Maruziyet eşit değil — ve az temsil edilenlere yükleniyor. Neredeyse kusursuz saldırıya en açık hastalar sistematik olarak veride **az temsil edilen gruplardan** geliyordu: etnik azınlıklar, nadir hastalık fenotipleri, alışılmadık görüntüleme özellikleri. Elektronik kayıt veri setinde Siyah hastalar en savunmasız kayıtlar arasında beklenenden **%31 daha sık** bulunuyordu; mamografi kümesinde malignite açısından şüpheli işaretlenmiş taramalar bu tehlike bölgesinde çarpıcı biçimde **+%1.179**

fazla temsil ediliyordu. (Bu iki rakam örnektir, bütün tablo değil — makale çeşitli gruplarda aynı kaymayı raporluyor — ve bunlar küçük, aşırı riskli kuyruktaki *göreli* aşırı temsildir: bu hastaların en açıkta kalan kayıtlar arasında ne kadar daha sık göründüğünü söyler; Siyah hastaların veya şüpheli mamografilerin yüzde kaçının açıkta olduğunu değil.) Modelin bu hastaları benzer örneklerin içinde kaybettirebileceği daha az örnek vardır; kayıtları öne çıkar — ve üyelik saldırısının yakaladığı şey tam olarak bu öne çıkmadır. Mahremiyet başarısızlığı rastgele değildir; zaten kenarda olan insanlarda yoğunlaşır.

Daha büyük modeller sorunu büyütüyor. Neredeyse kusursuz saldırılara açık hasta sayısı model kapasitesiyle keskin biçimde arttı — bir dermatoloji veri setinde pay, en küçük modelde neredeyse sıfırdan en büyükte yaklaşık **onda bire** çıktı. Tıbbi yapay zekâ modelleri büyüyüp daha yetenekli hale geldikçe — bütün alanın gittiği yön — yazarlar bu özel riskin azalmak yerine ağırlaştığı konusunda uyarıyor.

Rückert'in özeti keskin: "Bu kabul edilebilir bir risk değil. Sağlık verileri son derece hassastır."

"Ortalamada güvenli" neden yanlış testtir?

Düşük ortalama riski sorunsuz bir sağlık raporu gibi okumak cazip. Makalenin temel dersi, burada ortalama almanın yalnızca kaba değil — yanlış şeyi ölçüyor olması.

Bir mahremiyet garantisi ancak en fazla risk altındaki kişi için de geçerliyse anlamlıdır; ortadaki kişi için değil. 1.000 hastadan 999'unun üyeliğinin çıkarılamadığı ama birinin neredeyse kusursuz biçimde tanımlanabildiği bir model, o tek kişi açısından anlamlı herhangi bir biçimde "%99,9 mahrem" değildir — ve o kişi öngörülebilir biçimde nadir hastalığı olan veya etnik azınlıktan gelen hastaysa metrik yalnızca eksik değil, sessizce ayrımcıdır. Çoğunluğun güvenliğini ölçer ve buna herkesin güvenliği der.

Makalenin zorladığı değişim bu: *bu model ortalama olarak ne kadar mahrem?* sorusundan *en fazla açıkta olan hasta kim ve kimlerden biri?* sorusuna. Bunlar farklı sorular ve yalnızca ikincisi gerçek bir mahremiyet sorusu.

Bunun kanıtlamadığı — veya iddia etmediği — şeyler

- Tıbbi yapay zekânın **terk edilmesi gerektiğini söylemez**. Yazarların çerçevesi geri çekilmek değil, riski azaltmaktır: ölç, düzelt; geliştirmeyi bırakma.
- Her tıbbi modelin sızıntı yaptığı veya kullanımda olan herhangi bir modelin saldırıya uğradığı **anlamına gelmez**. *Riskin var olduğunu ve eşit dağılmadığını*, standart toplu metriklerin de bunu kaçırdığını gösterir.
- Kayıtlarınızın şu anda açıkta olduğu **anlamına gelmez**. Saldırı belirli koşullar gerektirir — modele erişim, aday kayıt ve saldırganın kendi altyapısı — sıradan, rastgele bir yetenek değildir.
- Hasta kaydının *içeriğinin* sızdığını **göstermez**. Sızan şey üyelik — dahil edilme gerçeği — ve bu

farklı bir nedenle tehlikelidir; kayıt dışarı döküldüğü için değil.

- Eşitsizlikleri tek bir manşet rakamına **indirgemez**. Güçlü ve tekrarlanabilir sonuç *örüntüdür*: toplu metrikler bireysel riski olduğundan düşük gösterir ve az temsil edilen hastalar bunun en büyük kısmını taşır — veri setleri ve yüzlerce model boyunca.

Kanıt ne kadar güçlü?

Bu deneysel ve metodolojik bir sonuç ve sağlam. Toplu mahremiyet metriklerinin hasta başına riski sistematik olarak olduğundan düşük göstermesi, kalan riskin az temsil edilen gruplarda yoğunlaşması ve model kapasitesiyle artması örüntüsü, **farklı veri türlerine sahip yedi veri setinde** ve her veri setinde çok sayıda modelde görüldü. Bir ölçüm iddiasını tek-veri-seti artefaktı olmaktan çıkarıp güvenilir yapan şey tam olarak bu genişlik.

İki dürüst uyarı var. Birincisi, bu yazarların kendi kurduğu saldırıları çalıştırarak ölçtüğü bir *risk* gösterimi; yetenekli bir saldırganın belirtilen varsayımlar altında ne kadar başarılı *olabileceğini* karakterize ediyor, gerçek dünyada saldırıların ne sıklıkta gerçekleştiğini değil. İkincisi, bazı hastalarda neredeyse kusursuz saldırı başarısı gibi çarpıcı sayılar tasarım gereği en kötü durumdaki bireyleri anlatıyor; bütün nokta bu ve "kuyruk ortalamanın ima ettiğinden çok daha ağır" diye okunmalı, "hastaların çoğu açıkta" diye değil.

Uygun tutum ne alarm ne de küçümseme: tıbbi yapay zekâ mahremiyetini sertifikalandırmak için kullandığımız standart yöntemin kendi en kötü vakalarını göremediğini ve bu vakaların en az hareket alanı olan hastalara düştüğünü gösteren dikkatli, çok-veri-setli bir çalışma.

Neden önemli?

Bunu teknik bir dipnot olmaktan çıkaran iki şey var.

Birincisi, "mahremiyeti koruyan" ifadesinin ne anlama gelmesine izin verilmesi gerektiğini değiştiriyor. Bir model ortalama mahremiyet skoruyla yayımlanırsa bu skor gerçekten düşük olabilir ve yine de bir üyelik saldırısıyla neredeyse kusursuz biçimde tanımlanabilen bir hasta alt kümesini gizleyebilir. Makalenin pratik talebi somut: yayımdan önce mahremiyet riskini **hasta başına** değerlendirin, kullanımdaki modellere kimin erişebileceğini kontrol edin ve **differential privacy** gibi teknikleri kullanın — eğitim sırasında üyelik saldırılarını körelten küçük, dikkatle ayarlanmış gürültü; modelin yararlılığına gerçek ve yönetilmesi gereken bir maliyet getirir (mahremiyet–yararlılık dengesi açıktır, bedava değildir). "Ortalamayı kontrol ettik" artık kontrol etmiş sayılmamalı.

İkincisi, mahremiyetle adaleti birbirine örüyor. Tıbbi verilerde az temsil edilen ve bu yüzden tıbbi yapay zekâ tarafından zaten daha kötü hizmet gören grupların, aynı yapay zekânın mahremiyetini de en az koruduğu ortaya çıkıyor. İlk eşitsizlik üzerinde yoğun çalışan bir alan ikincisini başkasının işi sayamaz. Aynı insanlar söz konusu.

Tıbbi yapay zekâ mahremiyetinin rahatlatıcı hikâyesi — *ölçtük, ortalama düşük, sorun yok* — bu makalenin söktüğü hikâye. İnsanları teknolojiden korkutmak için değil, standardı ait olduğu yere taşımak için: ancak zarar görme olasılığı en yüksek kişi için de kontrol ettiyseniz tuttuğunuzu iddia edebileceğiniz bir söz.

Kısa özet

Üyelik çıkarım saldırıları, belirli bir kişinin kaydının bir yapay zekâ modelinin eğitim verisinde olup olmadığını anlamaya çalışır — üyeliğin kendisi (örneğin belirli bir hastalığınız olduğunu) açığa çıkarabileceği için, alttaki kayıt hiç görünmese bile gerçek bir mahremiyet sızıntısıdır. Önceki çalışmalar bu saldırıların bir veri seti genelinde *ortalama* ne sıklıkta başarılı olduğunu ölçüyordu. Bu çalışma, yedi tıbbi veri setinde (görüntüleme, EKG, elektronik kayıtlar) ve her birinde birçok model üzerinde riski *hasta başına* ölçtü ve toplu metriklerin riski ciddi biçimde düşük gösterdiğini buldu: veri seti ortalaması güvenli görünse bile bazı bireyler neredeyse kusursuz biçimde tanımlanabiliyor (saldırı AUC $\geq$ 0,95); en savunmasızlar sistematik olarak az temsil edilen gruplardan (etnik azınlık, nadir hastalık, alışılmadık görüntüleme); sorun model büyüdükçe artıyor. Yazarlar tıbbi yapay zekâdan vazgeçilmesini savunmuyor — mahremiyetin birey düzeyinde ölçülmesini, model erişiminin kontrol edilmesini ve differential privacy kullanılmasını savunuyor. Sonuç: "ortalamada mahrem" bir mahremiyet garantisi değildir ve başarısız olduğu insanlar zaten en az korunanlardır.

Abartısız değerlendirme

Makale ne gösteriyor: Yedi tıbbi veri setinde ve her birinde çok sayıda modelde, hasta başına üyelik-çıkarım riski bazı bireylerde toplu metriklerin ima ettiğinden çok daha yüksek — neredeyse kusursuz saldırı başarısına (AUC $\geq$ 0,95) kadar — ve kalan risk az temsil edilen gruplara orantısız biçimde düşüyor, model kapasitesiyle de artıyor.

Makul ama kanıtlanmamış olan: Gerçek dünyadaki saldırganların bunu kullanımda olan klinik modellere karşı zaten kullandığı; çalışma belirtilen varsayımlar altında *yeteneği ve dağılımını* gösteriyor, sahadaki saldırı sıklığını değil.

Göstermediği şey: Tıbbi yapay zekânın terk edilmesi gerektiği; her modelin sızıntı yaptığı veya belirli bir kullanım modelinin ihlal edildiği; hasta kayıtlarının *içeriğinin* (üyelik yerine) açığa çıktığı; tek bir sayıyla özetlenebilecek eşitsizlik — sağlam sonuç tek sayı değil, örüntü.

Başlıca sınırlamalar: Gözlenmiş olayları değil, yazarların kendi saldırılarıyla en kötü durum riskini ölçüyor; çarpıcı sayılar tasarım gereği en açıkta olan bireyleri anlatıyor; grup farklarının kesin büyüklüğü tek sayıya indirgenmek yerine veri setleri boyunca tutarlı bir örüntü olarak gösteriliyor.

Genel okuyucu ne kadar güvenmeli? Toplu mahremiyet metriklerinin bireysel riski olduğundan düşük gösterdiğine, kalan riskin az temsil edilen hastalarda yoğunlaştığına ve bunun model boyutuyla kötüleştiğine yüksek güven — çok sayıda veri seti ve modelde gösterildi. Çalışmanın

ölçmediği gerçek dünya saldırı sıklığına orta güven. Güvenli okuma: ne “tıbbi AI verilerinizi sızdırıyor”, ne de “mahremiyet çözüldü”; daha doğrusu “standart mahremiyet kontrolü kendi en kötü vakalarını göremiyor ve bu vakalar en savunmasız hastalara düşüyor — dolayısıyla kontrolün değişmesi gerekiyor.”

Kaynaklar

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