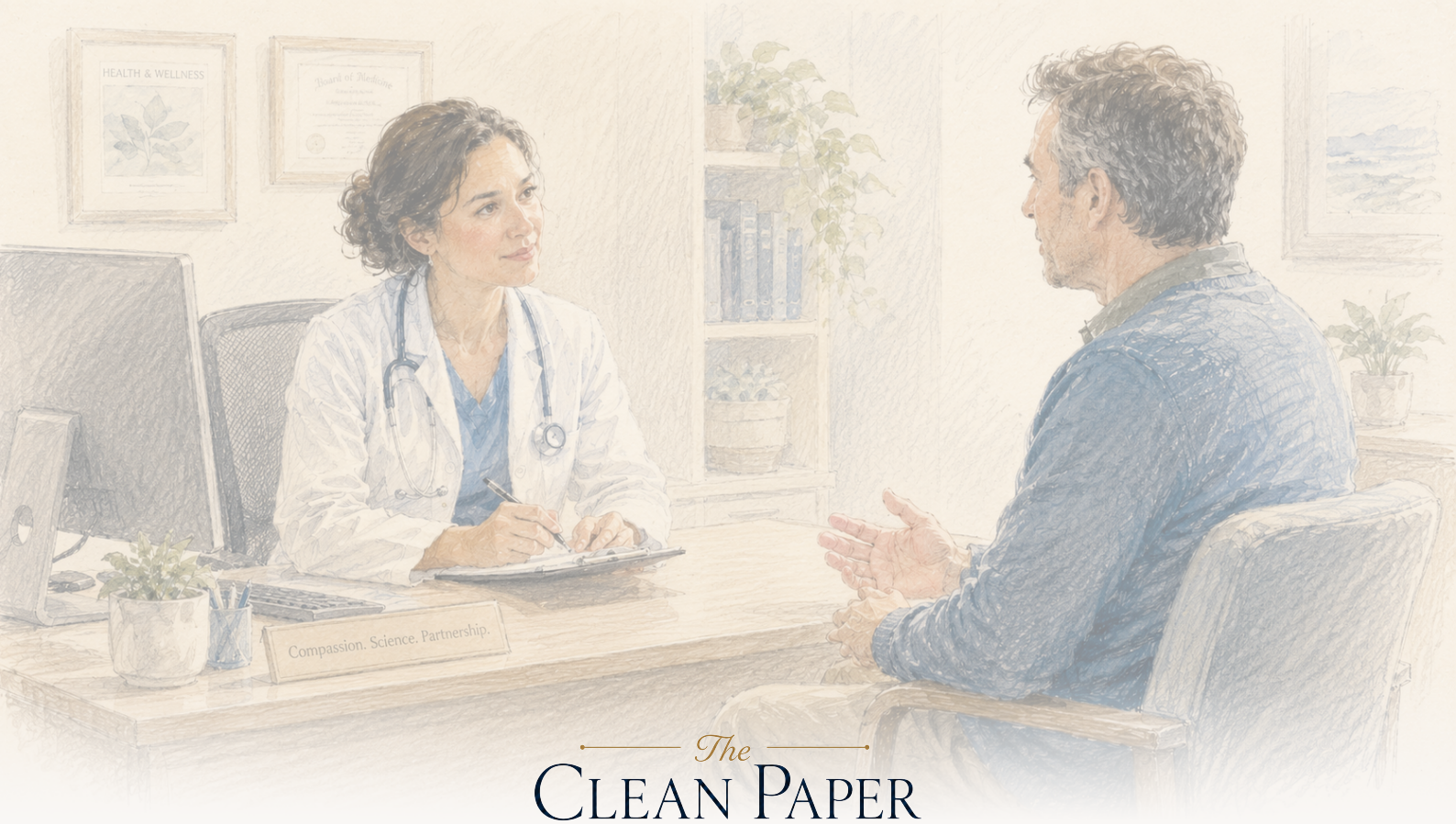


The Clean Paper

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

Güvenli görünen ve evrak işlerini iyileştiren klinik yapay zekâ hastaların sonuçlarını iyileştirmedi

Kenya'daki 16 birinci basamak sağlık tesisinde yürütülen pragmatik, küme-randomize bir çalışma, klinik görevlilerin kullandığı kayda eklenen üretken yapay zekâ destekli bir karar destek aracını test etti. 9,691 hasta arasında, tedavi başarısızlığının katı ve uzmanlarca karara bağlanan 14 günlük bileşik sonucu, aracı kullananlarda %2.2; kullanmayanlarda %2.0 idi (düzeltilmiş olasılık oranı 0.77, 95% CI 0.55-1.08, P = 0.13) — anlamlı bir fark yoktu. Araç herhangi bir güvenlik sinyali göstermedi, değerlendirilen tüm alanlarda dokümantasyonu iyileştirdi, reçetelemeyi ve memnuniyeti değiştirmedi ve antibiyotik maliyetlerinin düşmesi yönünde bir eğilim gösterdi. Bu başarısız bir çalışma değil; kıyaslama ölçütleriyle değil hastalar üzerinde ölçülmüş, nadir ve dürüst bir çalışma — güvenli görünen ve sürece yardımcı olan bir aracın iki hafta içinde hastalara ölçülebilir biçimde yardımcı olmadığı ve böyle bir yararı saptamanın çok daha büyük bir çalışma gerektireceği yönündeki gösterişsiz gerçeğe ulaşıyor.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial*, Ambrose Agweyu, Paul Mwaniki, Vaishnavi Menon, Robert Korom, Lynda Isaaka, Conrad Wanyama, Xiaoxuan Liu & Bilal A. Mateen (and colleagues), Nature Medicine (2026) · DOI: 10.1038/s41591-026-04503-6

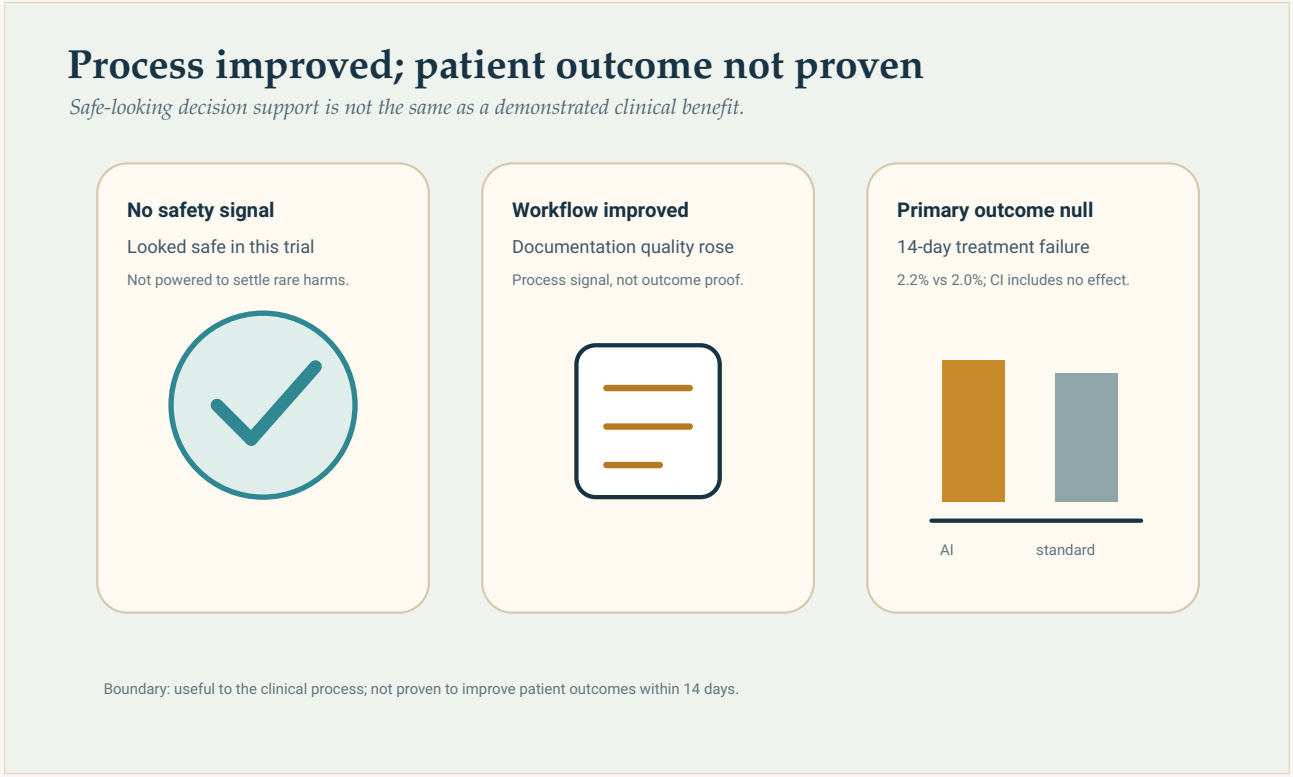
Güvenli ve yararlı olmak, etkili olmakla aynı şey değildir

Tıbbi yapay zekâ hakkındaki manşetlerin çoğu yanlış türdeki çalışmalara dayanıyor. Bir model sınav sorularında iyi puan alıyor ya da özenle seçilmiş klinik örneklerde doktorları geçiyor; ardından hikâye kendiliğinden yazılıyor: Makine klinik kullanıma hazır.

Bu çalışma daha zor ve daha nadir bir şey yaptı. Üretken yapay zekâ destekli bir karar destek aracını gerçek birinci basamak sağlık hizmetlerine, gerçek sağlık tesislerinde ve gerçek hastalarla birlikte yerleştirdi ve gerçekten önemli olan soruyu sordu: Hastaların sonuçları iyileşti mi?

Dürüst yanıt hayır — ölçülebilir biçimde değil, 14 gün içinde değil. Araç herhangi bir güvenlik sinyali göstermedi. Klinik dokümantasyonun kalitesini iyileştirdi. Hatta bazı ilaç maliyetlerini düşürmüş olabilir. Ancak tedavi başarısızlıklarını anlamlı biçimde azaltmadı ve yazarlar, herhangi bir yarar varsa bunun muhtemelen mütevazı olduğunu söylemekte dikkatli davranıyor.

Bu, çalışmanın başarısızlığı değil. Çalışmanın işini yapması. Yapay zekâya ilişkin sorumlu tıbbi kanıt, kıyaslama ölçütleri yerine hastalar üzerinde ölçüldüğünde böyle görünür.



Araç herhangi bir güvenlik sinyali göstermedi ve klinik dokümantasyonu iyileştirdi; ancak önceden belirlenmiş 14 günlük hasta sonucu anlamlı biçimde iyileşmedi. Sürece yardımcı olmak, hastaya sağladığı yararın kanıtlanmış olmasıyla aynı şey değildir.

The Clean Paper CC BY 4.0

Yazarlar ne yaptı?

Ekip, Kenya'nın Nairobi ve Kiambu ilçelerinde özel bir sağlık ağı (Penda Health) tarafından işletilen **16 birinci basamak sağlık tesisinde** pragmatik, küme-randomize bir çalışma yürüttü. Bu tesislerde sağlık hizmeti büyük ölçüde **klinik görevliler** tarafından sunuluyor — üç yıllık diplomaya sahip orta kademe sağlık çalışanları; bu kişiler çoğu zaman kıdemli meslektaşlarına kolayca danışamıyor.

Randomizasyon birimi hasta değil, klinisyendi. **103 klinik görevli** randomize edildi: 52'si müdahale koluna, 51'i kontrol koluna ayrıldı. Her iki kol da aynı bulut tabanlı elektronik tıbbi kayıt sistemini kullandı. Müdahale kolunda buna ek olarak, **OpenAI'nin GPT-4o** büyük dil modeli üzerine kurulu ve bu sisteme gömülü bir karar destek aracı olan **"AI Consult"** (sürüm 2.0) vardı. Araç, klinisyenin kaydettiği bilgileri okuyabiliyor ve tanı ya da tedavi planıyla ilgili olası sorunları işaretleyebiliyordu. Klinisyenlerin özerkliği tamdı: Önerileri kabul edebilir, değiştirebilir veya yok sayabilirlerdi.

> **Hangi modeldi ve ayrıntılar neden önemli?**
>
> Araç, **OpenAI'nin GPT-4o** modelini (Mayıs 2025 sürümü) çalıştıran **AI Consult 2.0** idi; OpenAI'nin ticari API'si üzerinden, kurumsal lisans kapsamında ve düşük rastlantısallık ayarlarında (temperature 0.1) kullanıldı. Araç, özel olarak geliştirilmiş bir elektronik kayıt sistemi olan EasyClinic'in EMR'si içinde çalışıyor ve Kenya'nın ulusal tedavi kılavuzlarıyla uyumlu olacak şekilde yazılmış sistem istemleriyle yönlendiriliyordu; yazarlar tam talimat istemini yayımladı.
>
> Bu ayrıntıları neden açıkça belirtmek gerekiyor? Çünkü sonuç genel olarak *"tıpta LLM'ler"* hakkında değil; *tek bir spesifik sistem* — tek bir model sürümü, tek bir istem, tek bir kayıt sistemi ve tek bir ortam — hakkındadır. Yazarlar da aynı noktayı vurguluyor: Bulgularını *sabit bir yetenek tahmini değil, zamana bağlı bir kıyaslama* olarak nitelendiriyorlar. Daha yeni bir model, farklı bir istem ya da daha az dijitalleşmiş bir klinik, sonucun tamamını değiştirebilir.
>
> Bağımsızlık konusunda: OpenAI daha sonra ayni destek (bulut bilişim kredileri ve API'sinin kullanımına ilişkin teknik rehberlik) sağladı; ancak yazarlar, OpenAI'yi kullanma kararının bu tekliften *önce* alındığını ve OpenAI'nin çalışmanın tasarımında, veri toplama sürecinde, analizinde ya da yayımlama kararında hiçbir rolünün olmadığını belirtiyor.

22 Nisan ile 16 Temmuz 2025 arasında **9,691 hasta** çalışmaya alındı. **Birincil sonuç kasıtlı olarak hasta merkezli ve katıydı**: ziyaretten sonraki 14 gün içinde ortaya çıkan ve uzmanlar tarafından karara bağlanan bir *tedavi başarısızlığı bileşik sonucu* — klinisyenlerden oluşan bir panel, çalışma kolunu bilmeden, her hastanın çözülemeyen ya da kötüleşen hastalık gibi kötü bir sonuç yaşayıp yaşamadığına karar verdi. Çalışma önceden kaydedilmişti (Pan-African Clinical Trials Registry 202502499779176).

Bu tasarım seçimi çalışmanın özüdür. Bir yapay zekâ aracının klinisyenin yazıya döktüğü şeyleri değiştirdiğini göstermek kolaydır. Hastanın başına gelenleri değiştirdiğini göstermek ise çok daha zor ve çok daha anlamlıdır.

Ne buldular?

Birincil sonuç iyileşmedi. Tedavi başarısızlığı, yapay zekâ kolunda 4,693 hastanın 102'sinde (%2.2), kontrol kolunda ise 4,654 hastanın 94'ünde (%2.0) görüldü. Ham yüzdeler yapay zekâ kolunda biraz daha yüksek olsa da düzeltme sonrasında nokta tahmini yarar yönündeydi: **düzeltilmiş olasılık oranı 0.77 (95% güven aralığı 0.55 ila 1.08, P = 0.13)** idi — istatistiksel olarak anlamlı değildi. Ham ve düzeltilmiş sayılar arasındaki bu farklılık bir aritmetik hatası değildir; düzeltme, klinisyen kümeleri arasındaki farklılıkları hesaba katar. Her iki durumda da güven aralığı rahatlıkla "etki yok" olasılığını içeriyor; dolayısıyla herhangi bir yarar ileri sürülemez ve mutlak açıdan etki çok küçüktü.

Olasılık oranlarının, güven aralıklarının ve P değerlerinin birlikte nasıl okunacağına ilişkin sade bir anlatım için klinik sonuçları okuma rehberine bakabilirsiniz.

Sınırları içinde herhangi bir güvenlik sinyali yoktu. Hiçbir ciddi advers olayın araçla ilişkili olduğu değerlendirilmedi ve bağımsız bir incelemede herhangi bir güvenlik sinyali saptanmadı. Yazarlar bu güvencenin sınırlarını dürüstçe ortaya koyuyor: Çalışma nadir görülen ciddi zararları saptamak üzere güçlendirilmemişti ve önceden belirlenmiş bir non-inferiority ya da resmî güvenlik çerçevesi yoktu; bu nedenle seyrek olaylar bakımından güvenliği *kanıtlayamaz*.

Dokümantasyon iyileşti. Körlenmiş uzmanlar tarafından incelenen 2,000 başvuru arasında, AI Consult kullanan klinisyenler değerlendirilen tüm alanlarda daha iyi klinik dokümantasyon üretti — kaydedilen tanı, tedavi planı ve genel bütünlük açısından.

Reçeteleme neredeyse değişmedi. Doğru antibiyotik kullanımı da dahil olmak üzere reçeteleme bakımından anlamlı bir fark yoktu (düzeltilmiş olasılık oranı 0.86, 95% CI 0.48 ila 1.55). Araç, antibiyotik reçeteleme oranlarını değiştirmedi.

Hastalar bir fark hissetmedi. Memnuniyet anketini tamamlayan 826 hasta arasında memnuniyet kollar arasında esasen aynıydı ve görüşme süreleri benzerdi.

Maliyetler hafifçe aşağı yönlüydü. Düzeltilmiş bir analizde, yapay zekâ kolunda antibiyotikle ilişkili maliyetler daha düşüktü — bunun muhtemel nedeni daha az reçete yazılması değil, daha ucuz seçeneklerin tercih edilmesiydi — ve hasta başına antibiyotik tasarrufu, aracı çalıştırmanın hasta başına maliyetini aşmış görünüyordu. Yazarlar bunu kesinleşmiş bir sonuçtan çok düşündürücü bir bulgu olarak sunuyor: toplam sahip olma maliyetinin tam muhasebesi çalışmanın kapsamı dışındaydı.

Yazarların kendi tek cümlelik özeti en açık biçimi sunuyor: LLM desteği bu sınırlar içinde güvenliydi, ancak 14 gün içinde tedavi başarısızlığını azaltmadı ve herhangi bir yarar muhtemelen mütevazıydı.

"Anlamlı fark yok" demek "işe yaramıyor" demek değildir

Birincil sonuçta sıfır bulgu, her iki yönde de kolayca aşırı yorumlanabilir. İki nokta basit anlatının önüne geçiyor.

İlk olarak, **çalışma gözlenenden daha büyük bir etkiyi saptayacak şekilde tasarlanmıştı.** Birinci basamak sağlık hizmetlerinde ciddi kötü sonuçlar nadirdir — burada yaklaşık %2 — bu nedenle küçük ama gerçek bir yararı gürültüden ayırmak çok büyük sayılar gerektirir. Yazarların kendi post-hoc güç hesaplamaları, gözlemledikleri büyüklükte bir etkiyi saptamak için **100,000'den fazla hasta ölçeğinde, çok daha büyük bir çalışma** gerekeceğini düşündürüyor. Bu büyüklükteki bir çalışmada anlamlı olmayan sonuç, küçük ama gerçek bir yararı dışlamaz; yalnızca bu çalışmanın böyle bir yararı kesinleştiremediği anlamına gelir.

İkinci olarak, **karşılaştırma kısmen bulanıktı.** Bir yapılandırma hatası, kısa bir süre boyunca bazı kontrol kolu klinisyenlerine AI Consult'a erişim sağladı ve ortak bir ağda çalışan klinisyenler birbirleriyle konuşuyor, sınırın ötesine alışkanlıklar taşıyor. Her iki etki de iki kolun birbirine daha çok benzemesine, gerçek bir farkın sıfıra doğru itilmesine yol açma eğilimindedir. Buna ek olarak, ev sahibi ağ zaten görece yüksek standartlarda çalışıyordu; bu da bir aracın iyileşme göstermesi için daha az alan bırakıyor.

Bunların hiçbiri "çığır açıcı" bir manşeti kurtarmaz. Ancak doğru okumanın küçümseyici değil, ölçülü olması gerektiği anlamına gelir: En zor ve en dürüst sonlanım noktasında bu araç, iki hafta içinde hastalara ölçülebilir biçimde yardımcı olmadı — buna karşılık kayıt tutma sürecine ölçülebilir biçimde yardımcı oldu ve güvenli göründü.

Bu çalışma neyi kanıtlamaz?

- Yapay zekânın hasta sonuçlarını iyileştirdiğini **göstermez**. Birincil 14 günlük sonlanım noktasında anlamlı bir yarar yoktu.
- Yapay zekânın işe yaramaz olduğunu **göstermez**. Nokta tahmini yapay zekâ lehindeydi, dokümantasyon her alanda iyileşti ve ilaç maliyetleri düşme eğilimi gösterdi; sıfır sonucu, çalışmanın doğrulamak için çok küçük kaldığı küçük bir gerçek yararla uyumludur.
- Aracın nadir görülen zararlar bakımından güvenli olduğunu **kanıtlamaz**. Herhangi bir güvenlik sinyali göstermedi, ancak nadir görülen ciddi olaylar bakımından güvenliği belgelemek üzere güçlendirilmemiş veya tasarlanmamıştı.
- "Yapay zekâ doktorları geçer" ya da klinisyenlerin yerini alır sonucunu **göstermez**. Bu bir karar *destek* aracıdır; klinisyen bunu kabul etme ya da reddetme konusunda tam yetkisini korudu.
- Sonuçları otomatik olarak genellemez. Çalışma Kenya'daki tek bir özel kentsel ağda yürütüldü; kırsal, kent çeperi ve yüksek gelirli ortamlar her iki yönde de farklılık gösterebilir.
- Maliyet tasarruflarını **kanıtlamaz**. Maliyet sinyali düşündürücüdür; tamamlanmış bir ekonomik değerlendirme değildir.

Kanıt ne kadar güçlü?

Temel iddia — kısa vadeli hasta zararında kanıtlanmış bir azalma olmadığı — bakımından kanıt, *tasarım açısından* güçlü ve *sonuç açısından* uygun ölçüde temkinlidir. Körlenmiş, hasta düzeyinde bileşik bir sonuç kullanan, ileriye dönük, önceden kaydedilmiş ve küme-randomize bir çalışma,

bunun gibi bir araç için toplanabilecek en iyi gerçek dünya kanıtına oldukça yakındır. Bir kıyaslama puanından ya da örnek olay çalışmasından çok daha bilgilendiricidir.

İkincil bulgular — daha iyi dokümantasyon, değişmeyen reçeteleme, benzer memnuniyet ve daha düşük antibiyotik maliyetleri — iyi kanıtlardır, ancak ikincil olarak okunmalıdır: başlık olacak sonuçlar değil, destekleyici sinyallerdir ve aynı bulaşma ile tek ağ sınırlamalarına açıktır.

Güvenlik açısından kanıt güven vericidir, ancak sınırlıdır: herhangi bir sinyal bulunmadı, fakat çalışma nadir görülen zararları bulmak üzere tasarlanmamıştı.

En yararlı tutum ne "işe yarıyor" ne de "başarısız oldu" demektir. Şudur: Dikkatle yürütülen gerçek dünya çalışması, bu yapay zekâ aracının herhangi bir güvenlik sinyali oluşturmadığını ve bakım sürecini iyileştirdiğini, ancak iki hafta içinde hasta sonuçlarında bir yarar gösteremediğini ortaya koydu — ve böyle bir yararı saptamak çok daha büyük bir çalışma gerektirirdi.

Neden önemli?

Tıbbi yapay zekâ tartışması tam da bu tür kanıtlardan yoksun. Modellerin sınavlarda başarılı olduğunu ve düzenli, temiz vakalarda klinisyenlerle eşleştiğini gösteren binlerce makale var. Gerçek hastaların durumunun daha iyi olup olmadığını ölçen büyük, pragmatik, randomize çalışmalar ise çok az. Bu çalışma onlardan biri ve gösterişsiz bir gerçeğe işaret ediyor: Sınavı geçmek, hastaya yardımcı olmakla aynı şey değildir.

Bütün hikâye bu boşlukta yatıyor. Bir araç klinisyenler için gerçekten yararlı olabilir — daha anlaşılır notlar, daha düşük ilaç faturaları, ikinci bir göz — ve yine de zor bir hasta sonucunu iki hafta içinde değiştirmeyebilir. Her iki gerçek aynı anda doğru olabilir ve olgun bir sağlık sistemi, işine geleni seçmek yerine bunları birlikte değerlendirmelidir.

Bu durum, kanıt yükünü de yararlı bir yönde yeniden belirliyor. Bir şirket klinik yapay zekâsının bakımı iyileştirdiğini iddia etmek istiyorsa, ilgili kanıt bir liderlik tablosu değildir. Hastalar için önemli sonuçlar üzerinde yürütülen bunun gibi bir çalışmadır — tercihen daha büyük bir çalışma; çünkü buradaki dürüst ders, mütevazı yararları görmek için büyük çalışmalara ihtiyaç olduğudur.

Kısa özet

Kenya'daki 16 birinci basamak sağlık tesisinde yürütülen pragmatik, küme-randomize bir çalışma, klinik görevlilerin kullandığı elektronik kayda eklenen üretken yapay zekâ destekli bir karar destek aracını ("AI Consult") test etti. 9,691 hasta arasında, 14 gün içindeki tedavi başarısızlığının uzmanlarca karara bağlanan bileşik sonucu, aracı kullananlarda %2.2; kullanmayanlarda %2.0 idi (düzeltilmiş olasılık oranı 0.77, 95% CI 0.55–1.08, P = 0.13) — anlamlı bir fark yoktu. Araç herhangi bir güvenlik sinyali göstermedi, değerlendirilen tüm alanlarda klinik dokümantasyonu iyileştirdi, reçetelemeyi değiştirmedi, hasta memnuniyetini aynı bıraktı ve biraz daha düşük antibiyotik maliyetleriyle ilişkiliydi. Yazarlar, bu sınırlar içinde güvenli olduğu, ancak tedavi başarısızlığını

azaltmadığı ve herhangi bir yararın muhtemelen mütevazı olduğu sonucuna varıyor; gözlenen büyüklükte bir etkiyi saptamak çok daha büyük bir çalışma (100,000 hasta ölçeğinde) gerektirirdi. Sonuç, klinik yapay zekânın hasta sonuçlarını iyileştirdiğini ya da işe yaramaz olduğunu göstermiyor — güvenli görünen ve sürece yardımcı olan bir aracın tek bir kentsel özel ağda iki hafta içinde hastalara ölçülebilir biçimde yardımcı olmadığını gösteriyor.

Abartısız değerlendirme

Makalenin gösterdiği: Gerçek dünya koşullarında yürütülen randomize bir çalışmada, birinci basamak sağlık kayıtlarına LLM tabanlı bir karar destek aracı eklenmesi herhangi bir güvenlik sinyali oluşturmadı, klinik dokümantasyon kalitesini iyileştirdi, reçetelemeyi değiştirmedi ve katı 14 günlük hasta düzeyindeki tedavi başarısızlığı bileşik sonucunu anlamlı biçimde azaltmadı.

Makul olan ancak kanıtlanmayan: Aracın, bu çalışmanın saptayamayacağı kadar küçük bir tedavi başarısızlığı azalması sağlaması; tam maliyetler hesaba katıldığında para tasarrufu sağlaması; daha iyi dokümantasyonun zamanla daha iyi bakıma dönüşmesi.

Gösterdiği şey değil: Klinik yapay zekânın hasta sonuçlarını iyileştirdiği; nadir olaylar bakımından güvenli olmadığı; klinisyenlerin yerini aldığı ya da onları geride bıraktığı; bu sonuçların kırsal ya da yüksek gelirli ortamlara aktarılabileceği; maliyet tasarrufunun kesinleştiği.

Başlıca sınırlamalar: Gözlemlenenden daha büyük bir etkiyi saptayacak şekilde güçlendirilmişti (nadir sonuçlar çok büyük örneklemler gerektirir); bir yapılandırma hatası kontrol kolunu kirletti ve ortak ağ alışkanlıkları karşılaştırmayı bulanıklaştırdı; bunların her ikisi de fark bulunmaması yönünde yanlılığa yol açar; hâlihazırda yüksek standartlara sahip tek bir özel kentsel ağ; önceden belirlenmiş bir non-inferiority veya güvenlik çerçevesinin bulunmaması; kısa 14 günlük zaman aralığı.

Genel bir okur ne kadar güven duymalı? Bu çalışmada aracın herhangi bir güvenlik sinyali göstermediği ve dokümantasyonu iyileştirdiği konusunda yüksek güven. Burada 14 günlük hasta sonuçlarını ölçülebilir biçimde iyileştirmediği konusunda yüksek güven. Hasta yararına yönelik bir müdahale olarak “işe yaradığı” ya da “başarısız olduğu” yönündeki her türlü iddia bakımından düşük güven — bu soru gerçekten çözümsüz durumda ve çok daha büyük bir çalışmaya ihtiyaç duyuyor. Uygun tutum şu: Birinci basamak sağlık hizmetlerini yapay zekânın dönüştürdüğü — ya da mahvettiği — yönünde bir hüküm değil, herhangi bir güvenlik sinyali oluşturmayan ve bakım sürecini iyileştiren bir araç hakkında dikkatle yürütülmüş, gerçek dünya sonucu.

Kaynaklar

Agweyu et al., Nature Medicine (2026)
https://doi.org/10.1038/s41591-026-04503-6

Pan-African Clinical Trials Registry, registration 202502499779176
https://pactr.samrc.ac.za/

EurekAlert / PATH press release on the trial (2026)
https://www.eurekalert.org/news-releases/1133583