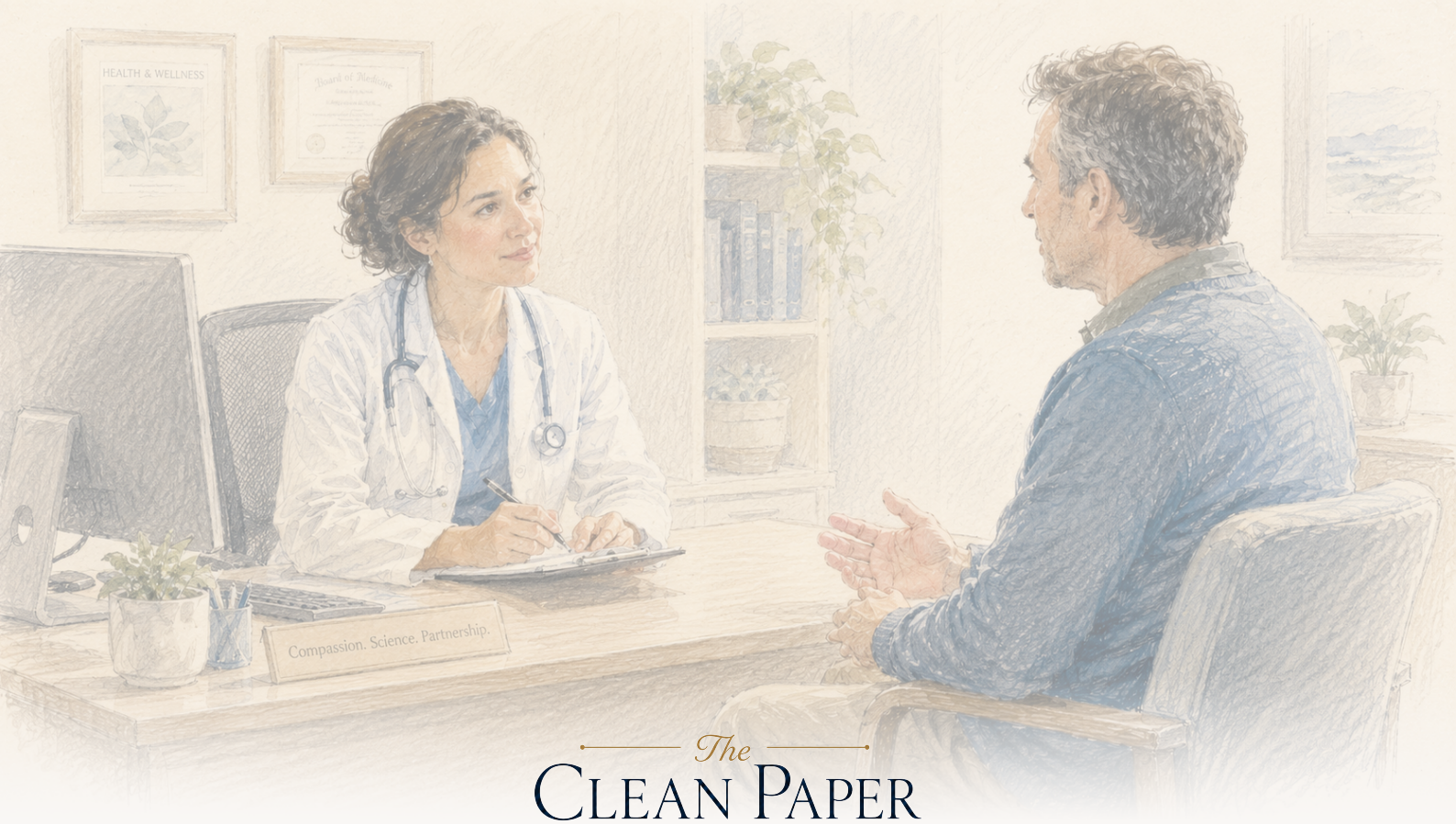


The Clean Paper

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

En klinisk AI som verkade säker och förbättrade pappersarbetet — men inte patientutfallen

En pragmatisk, klusterrandomiserad prövning vid 16 kenyanska primärvårdsinrättningar testade ett beslutsstödsverktyg med generativ AI som lagts till i den journal som användes av clinical officers. Bland 9 691 patienter var det strikta, expertbedömda sammansatta måttet på behandlingssvikt efter 14 dagar 2,2 % med verktyget mot 2,0 % utan det (justerad oddskvot 0,77, 95 % KI 0,55-1,08, P = 0,13) — ingen signifikant skillnad. Verktyget visade ingen säkerhetssignal, förbättrade dokumentationen inom samtliga bedömda områden, lämnade förskrivning och nöjdhet oförändrade och tenderade mot lägre antibiotikakostnader. Det är inte en misslyckad studie; det är en ovanlig, uppriktig studie — mätt på patienter, inte på riktmärken — och den landar i den oglamorösa sanningen att ett verktyg som verkade säkert och hjälpte processen inte bevisligen hjälpte patienter inom två veckor, och att det skulle krävas en betydligt större prövning för att upptäcka en eventuell sådan nytta.

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

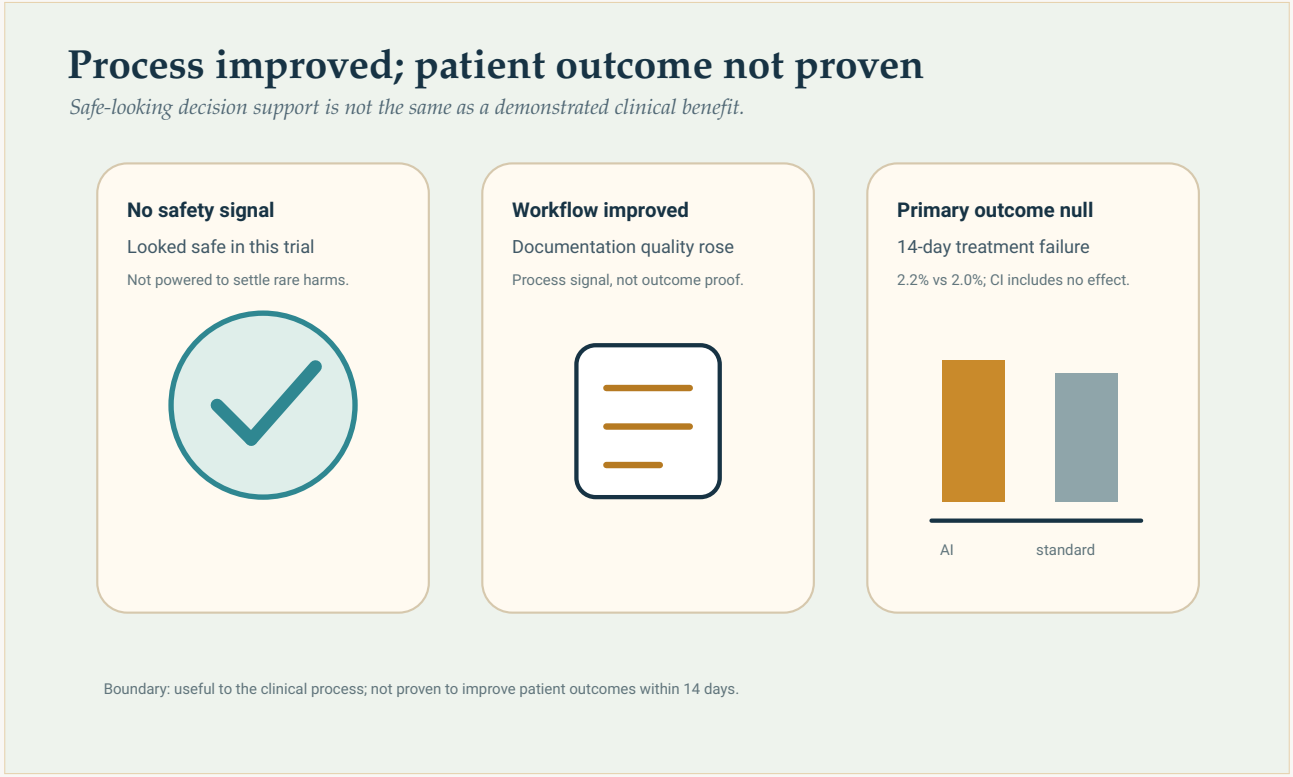
Säker och hjälpsam är inte samma sak som effektiv

De flesta rubriker om medicinsk AI bygger på fel sorts studie. En modell presterar bra på examensfrågor eller slår läkare på noggrant utvalda patientfall, och berättelsen skriver sig själv: maskinen är redo för kliniken.

Den här prövningen gjorde något svårare och ovanligare. Den satte in ett beslutsstödsverktyg med generativ AI i verklig primärvård, på verkliga vårdinrättningar, med verkliga patienter, och ställde frågan som faktiskt spelar roll: gick det bättre för patienterna?

Det ärliga svaret är nej — inte mätbart, inte inom 14 dagar. Verktyget visade ingen säkerhetssignal. Det förbättrade kvaliteten på den kliniska dokumentationen. Det kan till och med ha sänkt vissa läkemedelskostnader. Men det minskade inte behandlingssvikterna signifikant, och författarna är noga med att säga att en eventuell nytta sannolikt är blygsam.

Det är inte ett misslyckande för studien. Det är studien som fungerar. Så här ser ansvarsfull evidens om AI i medicinen ut när den mäts på patienter i stället för på riktmärken.



Verktyget visade ingen säkerhetssignal och förbättrade den kliniska dokumentationen, men det fördefinierade patientutfallet efter 14 dagar förbättrades inte signifikant. Hjälp i processen är inte samma sak som bevisad patientnytta.

Vad författarna gjorde

Forskargruppen genomförde en pragmatisk, klusterrandomiserad prövning vid **16 primärvårdsinrättningar** som drivs av ett privat vårdnätverk (Penda Health) i länen Nairobi och Kiambu i Kenya. Vården vid dessa inrättningar ges till stor del av **clinical officers** — vårdpersonal på mellannivå med en treårig diplomutbildning — ofta utan enkel tillgång till konsultation med en mer senior kollega.

Randomiseringsenheten var klinikern, inte patienten. **103 clinical officers** randomiserades: 52 till interventionsarmen och 51 till kontrollarmen. Båda armarna använde samma molnbaserade elektroniska patientjournal. Interventionsarmen hade dessutom **”AI Consult”** (version 2.0), ett beslutsstödsverktyg som byggde på den stora språkmodellen **OpenAI:s GPT-4o** och var integrerat i journalen. Det läste den information som en kliniker dokumenterade och kunde uppmärksamma möjliga problem med diagnosen eller behandlingsplanen. Klinikerna behöll full bestämmanderätt: de kunde godta, ändra eller bortse från förslagen.

> **Vilken modell var det, och varför spelar detaljerna roll**
>
> Verktyget var **AI Consult 2.0**, som körde **OpenAI:s GPT-4o** (utgåvan från maj 2025), nåddes via OpenAI:s kommersiella API med en företagslicens och kördes med inställningar för låg slumpmässighet (temperatur 0,1). Det var inbyggt i en skräddarsydd elektronisk journal (EasyClinics EMR) och styrdes av systempromptar som skrivits för att följa Kenyas nationella behandlingsriktlinjer; författarna publicerade hela instruktionsprompten.
> Varför ange allt detta? Därför att resultatet handlar om *ett specifikt system* — en modellversion, en prompt, en journal, en inställning — inte om ”stora språkmodeller i medicinen” i allmänhet. Författarna gör samma poäng: de kallar sitt fynd för ett *tidsbundet riktmärke snarare än en fast uppskattning av förmågan*. En nyare modell, en annan prompt eller en mindre digitaliserad klinik skulle alla kunna förändra utfallet.
> Om oberoendet: OpenAI bidrog senare med stöd in natura (krediter för molnberäkning och teknisk vägledning om användningen av dess API), men författarna uppger att beslutet att använda OpenAI fattades *före* erbjudandet och att OpenAI inte hade någon roll i prövningens utformning, datainsamling, analys eller beslutet att publicera.

Mellan den 22 april och den 16 juli 2025 inkluderades **9 691 patienter**. Det **primära utfallet var avsiktligt patientcentrerat och strikt**: ett expertbedömt *sammansatt mått på behandlingssvikt* inom 14 dagar efter besöket — en panel av kliniker, blindad för studiearm, bedömde om varje patient hade fått ett dåligt utfall, till exempel en sjukdom som inte hade gått över eller hade förvärrats. Prövningen registrerades i förväg (Pan-African Clinical Trials Registry 202502499779176).

Det designvalet är själva poängen. Det är lätt att visa att ett AI-verktyg förändrar vad en kliniker skriver ned. Det är mycket svårare, och mycket mer betydelsefullt, att visa att det förändrar vad som händer med patienten.

Vad de fann

Det primära utfallet förbättrades inte. Behandlingssvikt inträffade hos 102 av 4 693 patienter (2,2 %) i AI-armen och 94 av 4 654 (2,0 %) i kontrollarmen. De ojusterade procenttalen var marginellt högre med AI, men efter justering lutade punktskattningen åt nytta: den **justerade oddskvoten var 0,77 (95 % konfidensintervall 0,55 till 1,08, P = 0,13)** — inte statistiskt signifikant. Att riktningen vänder mellan de råa och justerade talen är inget räknefel; justeringen tar hänsyn till skillnader mellan klustren av kliniker. Oavsett vilket inkluderar konfidensintervallet med god marginal ”ingen effekt”, så ingen nytta kan hävdas, och i absoluta tal var effekten mycket liten.

För en lättillgänglig förklaring av hur oddskvoter, konfidensintervall och P-värden ska läsas tillsammans, se guiden till kliniska resultat.

Ingen säkerhetssignal — inom vissa gränser. Inga allvarliga biverkningar bedömdes vara relaterade till verktyget, och en oberoende granskning fann ingen säkerhetssignal. Författarna är öppna med gränsen för denna trygghet: prövningen saknade statistisk styrka för att upptäcka sällsynta allvarliga skador och hade inget fördefinierat ramverk för non-inferiority eller formell säkerhet, så den kan inte *bevisa* säkerhet vid ovanliga händelser.

Dokumentationen blev bättre. Bland 2 000 patientmöten som granskades av blindade experter producerade kliniker som använde AI Consult bättre klinisk dokumentation inom samtliga bedömda områden — den dokumenterade diagnosen, behandlingsplanen och den övergripande fullständigheten.

Förskrivningen förändrades knappt. Det fanns ingen signifikant skillnad i förskrivning, inklusive korrekt antibiotikaanvändning (justerad oddskvot 0,86, 95 % KI 0,48 till 1,55). Verktyget förändrade inte andelen antibiotikaförskrivningar.

Patienterna märkte ingen skillnad. Bland 826 patienter som besvarade en enkät om nöjdhet var nöjdheten i stort sett identisk mellan armarna, och besökstiderna var likartade.

Kostnaderna pekade svagt nedåt. I en justerad analys var de antibiotikarelaterade kostnaderna lägre i AI-armen — sannolikt genom billigare val snarare än färre förskrivningar — och besparingen på antibiotika per patient verkade överstiga kostnaden per patient för att driva verktyget. Författarna betecknar detta som en antydan, inte som avgjort: en fullständig beräkning av den totala ägandekostnaden låg utanför prövningen.

Författarnas egen sammanfattning i en mening är den tydligaste versionen: LLM-stödet var säkert inom dessa gränser men minskade inte behandlingssvikt inom 14 dagar, och en eventuell nytta är sannolikt blygsam.

Varför ”ingen signifikant skillnad” inte betyder ”det fungerar inte”

Ett nollresultat för det primära utfallet är lätt att övertolka i båda riktningarna. Två saker hindrar den enkla berättelsen.

För det första **utformades prövningen för att fånga en större effekt än den som hittades.** Allvarliga dåliga utfall i primärvården är sällsynta — omkring 2 % här — så det krävs enorma deltagarantal för att skilja en liten verklig nytta från brus. Författarnas egna post hoc-beräkningar av statistisk styrka antyder att det skulle krävas en **mycket större prövning, i storleksordningen mer än 100 000 patienter**, för att upptäcka en effekt av den storlek de observerade. Ett icke-signifikant resultat i en studie av den här storleken utesluter inte en liten, verklig nytta; det betyder att studien inte kunde avgöra om den fanns.

För det andra var **jämförelsen delvis sammanblandad.** Ett konfigurationsfel gav under en kort tid några kliniker i kontrollarmen tillgång till AI Consult, och kliniker i ett gemensamt nätverk talar med varandra och tar med sig vanor över gränsen. Båda effekterna tenderar att göra de två armarna mer lika och pressar därmed en eventuell verklig skillnad mot noll. Dessutom höll värdnätverket redan en relativt hög standard, vilket lämnar mindre utrymme för ett verktyg att visa en förbättring.

Inget av detta räddar en rubrik om ett ”genombrott”. Men det betyder att den korrekta tolkningen är nyanserad, inte avfärdande: på det svåraste och mest uppriktiga utfallsmåttet kunde det inte visas att verktyget hjälpte patienter på två veckor — samtidigt som det mätbart hjälpte journalföringen och verkade säkert.

Vad detta *inte* bevisar

- Det visar **inte** att AI:n förbättrade patientutfallen. För det primära utfallsmåttet efter 14 dagar fanns ingen signifikant nytta.
- Det visar **inte** att AI:n är oanvändbar. Punktskattningen talade till dess fördel, dokumentationen förbättrades genomgående och läkemedelskostnaderna tenderade att sjunka; nollresultatet är förenligt med en liten verklig nytta som prövningen var för liten för att bekräfta.
- Det bevisar **inte** att verktyget är säkert när det gäller sällsynta skador. Det visade ingen säkerhetssignal, men prövningen hade varken statistisk styrka eller en utformning som kunde fastställa säkerhet vid ovanliga allvarliga händelser.
- Det visar **inte** att ”AI slår läkare” eller ersätter kliniker. Detta är beslut*stöd*; klinikern behöll full befogenhet att godta eller avvisa det.
- Det kan **inte** automatiskt generaliseras. Prövningen genomfördes i ett enda privat urbant nätverk i Kenya; förhållanden på landsbygden, i stadsnära områden och i höginkomstmiljöer skulle kunna skilja sig i endera riktningen.
- Det fastställer **inte** kostnadsbesparingar. Kostnadssignalen är en antydan, inte en slutförd hälsoekonomisk utvärdering.

Hur starka är beläggen?

För det centrala påståendet — ingen bevisad minskning av kortsiktig patientskada — är evidensen stark *sett till designen* och lagom ödmjuk *sett till slutsatsen*. En prospektiv, förregistrerad, klusterrandomiserad prövning med ett blindat, sammansatt utfall på patientnivå ligger nära den bästa verklighetsbaserade evidens som går att samla in för ett sådant här verktyg. Den är långt mer informativ än ett riktmärkesresultat eller en studie med patientfall.

För de sekundära fynden — bättre dokumentation, oförändrad förskrivning, likartad nöjdhet, lägre antibiotikakostnader — är evidensen god men bör läsas som sekundär: stödjande signaler, inte huvudbudskapet, och sårbara för samma begränsningar med kontaminering och ett enda nätverk.

För säkerheten är evidensen betryggande men begränsad: ingen signal hittades, men studien var inte utformad för att hitta sällsynta skador.

Den mest användbara hållningen är varken ”det fungerar” eller ”det misslyckades”. Den är: en noggrann verklighetsnära prövning fann att det här AI-verktyget inte gav någon säkerhetssignal och förbättrade vårdprocessen, utan att visa någon nytta för patientutfallen inom två veckor — och att det skulle krävas en betydligt större studie för att upptäcka en eventuell sådan nytta.

Varför det spelar roll

Debatten om medicinsk AI lider brist på precis den här sortens evidens. Det finns tusentals artiklar som visar modeller som klarar prov galant och matchar kliniker i tillrättalagda fall. Det finns mycket få stora, pragmatiska, randomiserade prövningar som mäter om verkliga patienter får det bättre. Det här är en av dem, och den landar i den oglamorösa sanningen: att klara provet är inte samma sak som att hjälpa patienten.

Det glappet är hela berättelsen. Ett verktyg kan vara genuint användbart för kliniker — tydligare anteckningar, lägre läkemedelskostnader, ett extra par ögon — och ändå inte förändra ett krävande patientutfall på två veckor. Båda sakerna kan vara sanna samtidigt, och ett moget hälso- och sjukvårdssystem måste hålla ihop dem i stället för att välja den bekväma.

Det flyttar också bevisbördan i en användbar riktning. Om ett företag vill hävda att dess kliniska AI förbättrar vården är den relevanta evidensen inte en topplista. Det är en prövning som denna, med utfall som spelar roll för patienter — och helst en större prövning, eftersom den uppriktiga lärdomen här är att blygsamma nyttor kräver stora studier för att kunna upptäckas.

Kort sammanfattning

En pragmatisk, klusterrandomiserad prövning vid 16 kenyanska primärvårdsinrättningar testade ett beslutsstödsverktyg med generativ AI (”AI Consult”) som lagts till i den elektroniska journal som an-

vändes av clinical officers. Bland 9 691 patienter var det expertbedömda, sammansatta måttet på behandlingssvikt inom 14 dagar 2,2 % med verktyget mot 2,0 % utan det (justerad oddskvot 0,77, 95 % KI 0,55–1,08, P = 0,13) — ingen signifikant skillnad. Verktyget visade ingen säkerhetssignal, förbättrade den kliniska dokumentationen inom samtliga bedömda områden, förändrade inte förskrivningen, lämnade patientnöjdheten oförändrad och var förknippat med något lägre antibiotikakostnader. Författarna drar slutsatsen att det var säkert inom dessa gränser men inte minskade behandlingssvikt, och att en eventuell nytta sannolikt är blygsam; för att upptäcka en effekt av den observerade storleken skulle det krävas en mycket större prövning (i storleksordningen 100 000 patienter). Resultatet visar varken att klinisk AI förbättrar patientutfall eller att den är oanvändbar — det visar att ett verktyg som verkade säkert och hjälpte processen inte bevisligen hjälpte patienter inom två veckor, i ett enda privat urbant nätverk.

Rakt på sak

Vad artikeln visar: I en randomiserad prövning i verklig vård gav tillägget av ett LLM-baserat beslutsstödsverktyg i primärvårdens journaler ingen säkerhetssignal, förbättrade kvaliteten på den kliniska dokumentationen, förändrade inte förskrivningen och minskade inte signifikant ett strikt sammansatt mått på behandlingssvikt hos patienter efter 14 dagar.

Vad som är rimligt men inte bevisat: Att verktyget ger en liten verklig minskning av behandlingssvikt som var för liten för att prövningen skulle kunna upptäcka den; att det sparar pengar när samtliga kostnader räknas in; att bättre dokumentation med tiden leder till bättre vård.

Vad den inte visar: Att klinisk AI förbättrar patientutfall; att den är osäker vid sällsynta händelser; att den ersätter eller överträffar kliniker; att dessa resultat kan överföras till landsbygds- eller höginkomstmiljöer; att kostnadsbesparingen är fastställd.

Huvudsakliga begränsningar: Dimensionerad för en större effekt än den observerade (sällsynta utfall kräver mycket stora urval); ett konfigurationsfel kontaminerade kontrollarmen och vanor inom det gemensamma nätverket suddar ut jämförelsen, vilket i båda fallen förskjuter resultatet mot ingen skillnad; ett enda privat urbant nätverk med redan hög standard; inget fördefinierat ramverk för non-inferiority eller säkerhet; kort tidshorisont på 14 dagar.

Hur stort förtroende bör en allmän läsare ha? Stort förtroende för att verktyget inte visade någon säkerhetssignal i denna prövning och förbättrade dokumentationen. Stort förtroende för slutsatsen att det inte bevisligen förbättrade patientutfallen efter 14 dagar här. Litet förtroende för varje påstående om att det ”fungerar” eller ”misslyckas” som en intervention för patientnytta — den frågan är genuint olöst och kräver en mycket större studie. En lämplig hållning: ett noggrant, verklighetsnära resultat om ett verktyg som inte gav någon säkerhetssignal och förbättrade vårdprocessen, inte ett utslag om att AI omvandlar — eller förstör — primärvården.

Källor

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