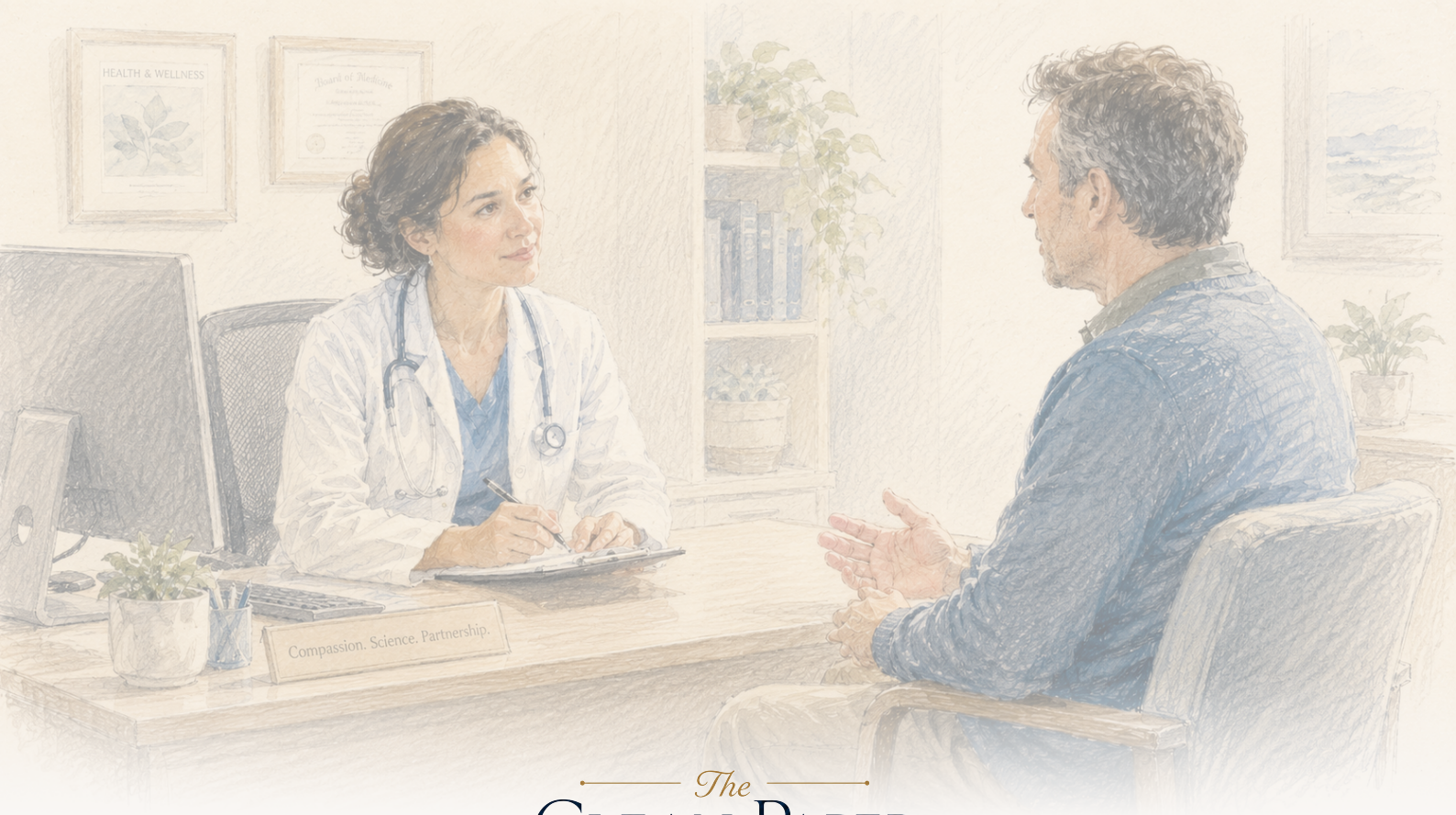


The Clean Paper

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

En klinisk KI som så trygg ut og forbedret papirarbeidet — men ikke pasientutfallene

En pragmatisk, klyngerandomisert utprøving ved 16 kenyanske primærhelseklinikker testet et beslutningsstøtteverktøy med generativ KI som var lagt til i journalen clinical officers brukte. Blant 9 691 pasienter var det strenge, ekspertvurderte, sammensatte målet på behandlingssvikt etter 14 dager 2,2 % med verktøyet mot 2,0 % uten (justert oddsratio 0,77, 95 % KI 0,55-1,08, P = 0,13) — ingen signifikant forskjell. Verktøyet viste ikke noe sikkerhetssignal, forbedret dokumentasjonen på alle vurderte områder, lot forskrivning og tilfredshet være uendret og viste en tendens til lavere antibiotikakostnader. Det er ikke en mislykket studie; det er en sjelden, ærlig studie — målt på pasienter, ikke på referansetester — og den lander på den lite glamorøse sannheten at et verktøy som så trygt ut og hjalp prosessen, ikke påviselig hjalp pasienter i løpet av to uker, og at det ville kreve en langt større utprøving å oppdage en eventuell slik nytte.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · 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

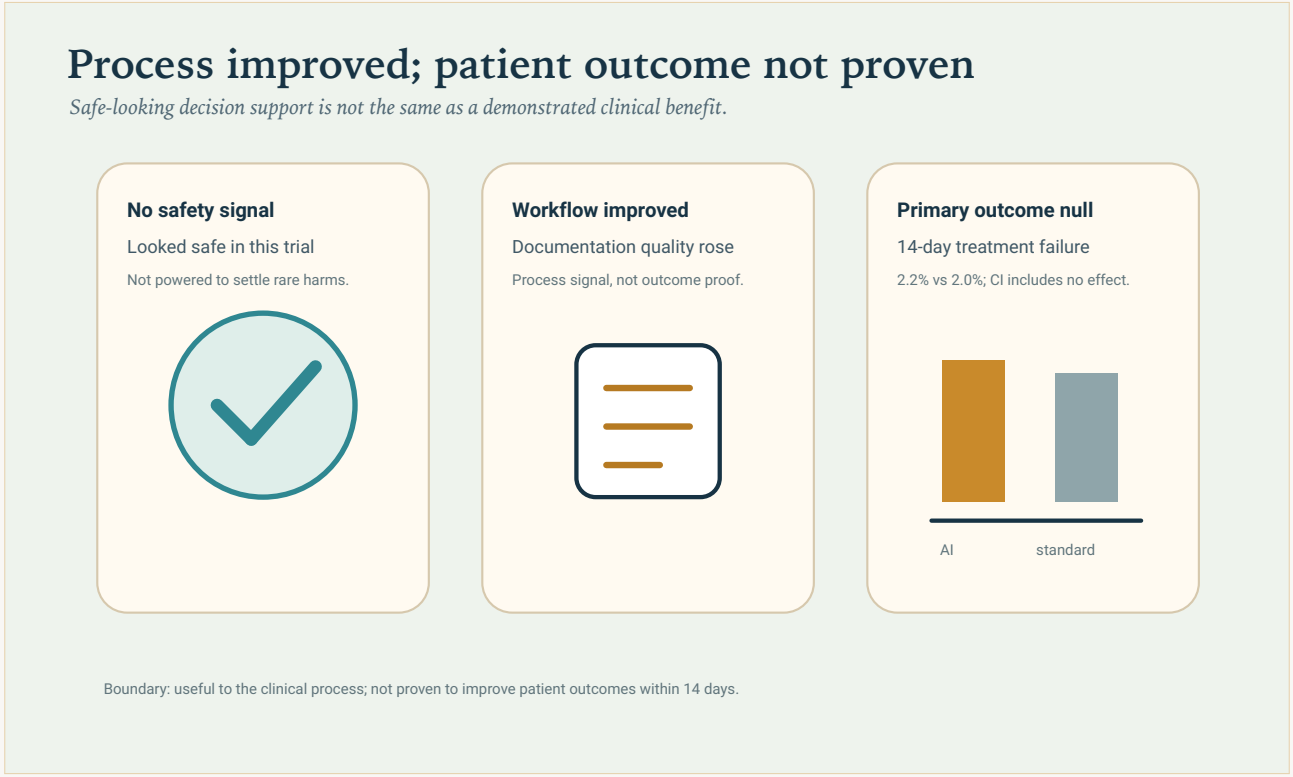
Trygt og nyttig er ikke det samme som effektivt

De fleste overskrifter om medisinsk KI bygger på feil type studie. En modell gjør det godt på eksamen-soppgaver eller slår leger på kuraterte pasienthistorier, og historien skriver seg selv: Maskinen er klar for klinikken.

Denne utprøvingen gjorde noe vanskeligere og sjeldnere. Den satte et beslutningsstøtteverktøy med generativ KI inn i reell primærhelsetjeneste, på virkelige klinikker, med virkelige pasienter, og stilte spørsmålet som faktisk betyr noe: Gikk det bedre med pasientene?

Det ærlige svaret er nei — ikke målbart, ikke innen 14 dager. Verktøyet viste ikke noe sikkerhetssignal. Det forbedret kvaliteten på den kliniske dokumentasjonen. Det kan til og med ha redusert enkelte legemiddelkostnader. Men det reduserte ikke behandlingssvikt signifikant, og forfatterne er nøye med å si at en eventuell nytte sannsynligvis er beskjeden.

Det er ikke en mislykket studie. Det er studien som virker. Slik ser ansvarlig evidens om KI i medisin ut når den måles på pasienter i stedet for på referansetester.



Verktøyet viste ikke noe sikkerhetssignal og forbedret den kliniske dokumentasjonen, men det forhåndsspesifiserte pasientutfallet etter 14 dager ble ikke signifikant bedre. Hjelp i prosessen er ikke det samme som dokumentert pasientnytte.

Hva forfatterne gjorde

Forskergruppen gjennomførte en pragmatisk, klyngerandomisert utprøving ved **16 primærhelseklinikker** drevet av et privat helsenettverk (Penda Health) i fylkene Nairobi og Kiambu i Kenya. Behandlingen ved disse klinikkene gis i stor grad av **clinical officers** — helsepersonell på mellomnivå med en treårig diplomutdanning — ofte uten enkel tilgang til konsultasjon med en mer erfaren kollega.

Randomiseringsenheten var klinikeren, ikke pasienten. **103 clinical officers** ble randomisert: 52 til intervensjonsarmen og 51 til kontrollarmen. Begge armene brukte den samme skybaserte elektroniske pasientjournalen. Intervensjonsarmen hadde i tillegg **«AI Consult»** (versjon 2.0), et beslutningsstøtteverktøy bygget på den store språkmodellen **OpenAIs GPT-4o** og integrert i journalen. Det leste informasjonen klinikeren dokumenterte, og kunne varsle om mulige problemer med diagnosen eller behandlingsplanen. Klinikerne beholdt full selvbestemmelse: De kunne godta, endre eller se bort fra forslagene.

> **Hvilken modell var det, og hvorfor detaljene betyr noe**
>
> Verktøyet var **AI Consult 2.0**, som kjørte **OpenAIs GPT-4o** (utgaven fra mai 2025), ble brukt gjennom OpenAIs kommersielle API med en bedriftslisens og kjørt med innstillinger for lav tilfeldighet (temperatur 0,1). Det var integrert i en skreddersydd elektronisk journal (EasyClinics EMR) og styrt av systeminstrukser skrevet for å samsvare med Kenyas nasjonale retningslinjer for behandling; forfatterne publiserte hele instruksjonsprompten.
> Hvorfor presisere alt dette? Fordi resultatet gjelder *ett bestemt system* — én modellversjon, én prompt, én journal, én innstilling — ikke «store språkmodeller i medisin» generelt. Forfatterne understreker det samme: De kaller funnet sitt en *tidsbestemt referanseverdi snarere enn et fast estimat av kapasiteten*. En nyere modell, en annen prompt eller en mindre digitalisert klinikk kan alle endre utfallet.
> Om uavhengighet: OpenAI ga senere støtte i form av naturalytelser (kreditter til nettskyberegning og teknisk veiledning i bruk av selskapets API), men forfatterne oppgir at beslutningen om å bruke OpenAI ble tatt *før* dette tilbudet, og at OpenAI ikke hadde noen rolle i utformingen av utprøvingen, datainnsamlingen, analysen eller beslutningen om å publisere.

Mellom 22. april og 16. juli 2025 ble **9 691 pasienter** inkludert. Det **primære utfallet var bevisst pasientsentrert og strengt**: et ekspertvurdert *sammensatt mål på behandlingssvikt* innen 14 dager etter konsultasjonen — et panel av klinikere, blindet for studiearm, vurderte om hver pasient hadde hatt et dårlig utfall, for eksempel en sykdom som ikke hadde gått over eller hadde blitt verre. Utprøvingen var registrert på forhånd (Pan-African Clinical Trials Registry 202502499779176).

Dette designvalget er selve poenget. Det er lett å vise at et KI-verktøy endrer det en kliniker skriver ned. Det er langt vanskeligere, og langt mer meningsfullt, å vise at det endrer det som skjer med pasienten.

Hva de fant

Det primære utfallet ble ikke bedre. Behandlingssvikt oppsto hos 102 av 4 693 pasienter (2,2 %) i KI-armen og 94 av 4 654 (2,0 %) i kontrollarmen. De ujusterte prosentandelene var marginalt høyere med KI, men etter justering pekte punktestimatet i retning av nytte: **justert oddsratio var 0,77 (95 % konfidensintervall 0,55 til 1,08, P = 0,13)** — ikke statistisk signifikant. At retningen skifter mellom de rå og justerte tallene, er ikke en regnefeil; justeringen tar hensyn til forskjeller mellom klinikerklyngene. Uansett inneholder konfidensintervallet med god margin «ingen effekt», så ingen nytte kan hevdes, og i absolutte tall var effekten svært liten.

For en lettfattelig forklaring på hvordan oddsratioer, konfidensintervaller og P-verdier skal leses sammen, se veiledningen til kliniske resultater.

Ikke noe sikkerhetssignal — innenfor visse grenser. Ingen alvorlige bivirkninger ble vurdert som knyttet til verktøyet, og en uavhengig gjennomgang fant ikke noe sikkerhetssignal. Forfatterne er ærlige om grensen for denne betryggelsen: Utprøvingen hadde ikke statistisk styrke til å oppdage sjeldne, alvorlige skader, og den hadde ikke noe forhåndsspesifisert rammeverk for non-inferioritet eller formell sikkerhet. Den kan derfor ikke *bevise* sikkerhet ved uvanlige hendelser.

Dokumentasjonen ble bedre. Blant 2 000 konsultasjoner gjennomgått av blindede eksperter produserte klinikere som brukte AI Consult, bedre klinisk dokumentasjon på alle vurderte områder — den dokumenterte diagnosen, behandlingsplanen og den samlede fullstendigheten.

Forskrivningen endret seg knapt. Det var ingen signifikant forskjell i forskrivning, heller ikke i riktig antibiotikabruk (justert oddsratio 0,86, 95 % KI 0,48 til 1,55). Verktøyet endret ikke andelen antibiotikaforskrivninger.

Pasientene merket ingen forskjell. Blant 826 pasienter som fullførte en tilfredshetsundersøkelse, var tilfredsheten i hovedsak identisk mellom armene, og konsultasjonstidene var omtrent like.

Kostnadene pekte svakt nedover. I en justert analyse var de antibiotikarelaterte kostnadene lavere i KI-armen — trolig gjennom billigere valg snarere enn færre forskrivninger — og besparelsen på antibiotika per pasient så ut til å overstige kostnaden per pasient ved å drive verktøyet. Forfatterne omtaler dette som et antydende, ikke avklart funn: En fullstendig beregning av totale eierkostnader lå utenfor utprøvingen.

Forfatternes egen oppsummering i én setning er den tydeligste versjonen: LLM-støtten var trygg innenfor disse grensene, men reduserte ikke behandlingssvikt innen 14 dager, og en eventuell nytte er sannsynligvis beskjeden.

Hvorfor «ingen signifikant forskjell» ikke betyr «det virker ikke»

Et nullresultat for det primære utfallet er lett å overtolke i begge retninger. To ting hindrer den enkle fortellingen.

For det første var **utprøvingen laget for å fange opp en større effekt enn den fant.** Alvorlige dårlige utfall i primærhelsetjenesten er sjeldne — rundt 2 % her — så det krever enorme deltakerantall å skille en liten, reell nytte fra støy. Forfatternes egne post hoc-beregninger av statistisk styrke tyder på at det ville kreve en **langt større utprøving, i størrelsesorden mer enn 100 000 pasienter**, å oppdage en effekt av den størrelsen de observerte. Et ikke-signifikant resultat i en studie av denne størrelsen utelukker ikke en liten, reell nytte; det betyr at denne studien ikke kunne avgjøre om den fantes.

For det andre var **sammenligningen delvis utvisket.** En konfigurasjonsfeil ga noen klinikere i kontrollarmen kortvarig tilgang til AI Consult, og klinikere i et felles nettverk snakker med hverandre og tar med seg vaner over grensen. Begge deler gjør gjerne de to armene mer like og skyver dermed en eventuell reell forskjell mot null. I tillegg holdt vertsnettverket allerede en relativt høy standard, noe som gir et verktøy mindre rom til å vise forbedring.

Ikke noe av dette redder en overskrift om et «gjennombrudd». Men det betyr at riktig tolkning er avstemt, ikke avvisende: På det vanskeligste og mest ærlige endepunktet kunne det ikke påvises at dette verktøyet hjalp pasienter i løpet av to uker — samtidig som det målbart hjalp journalføringen og så trygt ut.

Hva dette *ikke* beviser

- Det viser **ikke** at KI-en forbedret pasientutfallene. På det primære endepunktet etter 14 dager var det ingen signifikant nytte.
- Det viser **ikke** at KI-en er unyttig. Punktestimatet talte i dens favør, dokumentasjonen ble bedre over hele linjen, og legemiddelkostnadene viste en nedadgående tendens; nullresultatet er forenlig med en liten, reell nytte som utprøvingen var for liten til å bekrefte.
- Det beviser **ikke** at verktøyet er trygt med tanke på sjeldne skader. Det viste ikke noe sikkerhetssignal, men utprøvingen hadde verken statistisk styrke eller en utforming som kunne dokumentere sikkerhet ved uvanlige, alvorlige hendelser.
- Det viser **ikke** at «KI slår leger» eller erstatter klinikere. Dette er beslutnings*støtte*; klinikeren beholdt full myndighet til å godta eller avvise den.
- Det kan **ikke** generaliseres automatisk. Utprøvingen ble gjennomført i ett privat, urbant nettverk i Kenya; forhold på landsbygda, i bynære områder og i høyinntektsmiljøer kan avvike i begge retninger.
- Det fastslår **ikke** kostnadsbesparelser. Kostnadssignalet er antydende, ikke en fullført helseøkonomisk evaluering.

Hvor sterk er evidensen?

For den sentrale påstanden — ingen dokumentert reduksjon i kortsiktig pasientskade — er evidensen sterk *som design* og passende nøktern *som konklusjon*. En prospektiv, forhåndsregistrert, klyngerandomisert utprøving med et blindet, sammensatt utfall på pasientnivå er nær den beste evidensen fra den virkelige verden man kan samle inn for et slikt verktøy. Den er langt mer informativ enn et resultat på en referansetest eller en studie av pasienthistorier.

For de sekundære funnene — bedre dokumentasjon, uendret forskrivning, lik tilfredshet, lavere antibiotikakostnader — er evidensen god, men bør leses som sekundær: støttende signaler, ikke hovedbudskapet, og utsatt for de samme begrensningene knyttet til kontaminering og ett enkelt nettverk.

For sikkerhet er evidensen betryggende, men avgrenset: Ingen signal ble funnet, men dette var ikke en studie laget for å finne sjeldne skader.

Den mest nyttige holdningen er verken «det virker» eller «det mislyktes». Den er: En grundig utprøving i den virkelige verden fant at dette KI-verktøyet ikke ga noe sikkerhetssignal og forbedret behandlingsprosessen, uten å dokumentere noen nytte for pasientutfallene i løpet av to uker — og at det ville kreve en langt større studie å oppdage en eventuell slik nytte.

Hvorfor det er viktig

Debatten om medisinsk KI mangler nettopp denne typen evidens. Det finnes tusenvis av artikler som viser modeller som briljerer på eksamener og matcher klinikere i ryddige kasus. Det finnes svært få store, pragmatiske, randomiserte utprøvinger som måler om virkelige pasienter får det bedre. Dette er en av dem, og den lander på den lite glamorøse sannheten: Å bestå testen er ikke det samme som å hjelpe pasienten.

Dette gapet er hele historien. Et verktøy kan være genuint nyttig for klinikere — tydeligere notater, lavere legemiddelregninger, et ekstra par øyne — og likevel ikke flytte et krevende pasientutfall på to uker. Begge deler kan være sant samtidig, og et modent helsevesen må holde dem sammen i stedet for å velge det som passer best.

Det flytter også bevisbyrden i en nyttig retning. Hvis et selskap vil hevde at deres kliniske KI forbedrer behandlingen, er den relevante evidensen ikke en resultatliste. Det er en utprøving som denne, på utfall som betyr noe for pasienter — og helst en større en, fordi den ærlige lærdommen her er at beskjedne nytteeffekter krever store studier for å bli oppdaget.

Kort oppsummering

En pragmatisk, klyngerandomisert utprøving ved 16 kenyanske primærhelseklinikker testet et beslutningsstøtteverktøy med generativ KI («AI Consult») som var lagt til i den elektroniske journalen clinical officers brukte. Blant 9 691 pasienter var det ekspertvurderte, sammensatte målet på behandlingssvikt innen 14 dager 2,2 % med verktøyet mot 2,0 % uten (justert oddsratio 0,77, 95 % KI 0,55–1,08, P = 0,13) — ingen signifikant forskjell. Verktøyet viste ikke noe sikkerhetssignal, forbedret den kliniske dokumentasjonen på alle vurderte områder, endret ikke forskrivningen, lot pasienttilfredsheten være uendret og var forbundet med noe lavere antibiotikakostnader. Forfatterne konkluderer med at det var trygt innenfor disse grensene, men ikke reduserte behandlingssvikt, og at en eventuell nytte sannsynligvis er beskjeden. For å oppdage en effekt av den observerte størrelsen ville det kreves en langt større utprøving (i størrelsesorden 100 000 pasienter). Resultatet viser verken at klinisk KI forbedrer pasientutfall, eller at den er unyttig — det viser at et verktøy som så trygt ut og hjalp prosessen, ikke påviselig hjalp pasienter i løpet av to uker, i ett privat, urbant nettverk.

No-BS-sjekk

Hva artikkelen viser: I en randomisert utprøving i den virkelige verden ga tilføyelsen av et LLM-basert beslutningsstøtteverktøy i pasientjournalene i primærhelsetjenesten ikke noe sikkerhetssignal, forbedret kvaliteten på den kliniske dokumentasjonen, endret ikke forskrivningen og reduserte ikke signifikant et strengt, sammensatt mål på behandlingssvikt hos pasienter etter 14 dager.

Hva som er plausibelt, men ikke bevist: At verktøyet gir en liten, reell reduksjon i behandlingssvikt som var for liten til at denne utprøvingen kunne oppdage den; at det sparer penger når alle kostnader regnes med; at bedre dokumentasjon etter hvert gir bedre behandling.

Hva den ikke viser: At klinisk KI forbedrer pasientutfall; at den er utrygg ved sjeldne hendelser; at den erstatter eller overgår klinikere; at disse resultatene kan overføres til landsbygda eller høyinntektsmiljøer; at kostnadsbesparelsen er fastslått.

Viktigste begrensninger: Dimensjonert for en større effekt enn den observerte (sjeldne utfall krever svært store utvalg); en konfigurasjonsfeil kontaminerte kontrollarmen, og vaner i det felles nettverket visker ut sammenligningen, noe som i begge tilfeller trekker resultatet mot ingen forskjell; ett privat, urbant nettverk med allerede høy standard; ikke noe forhåndsspesifisert rammeverk for non-inferioritet eller sikkerhet; kort tidshorisont på 14 dager.

Hvor stor tillit bør en vanlig leser ha? Høy tillit til at verktøyet ikke viste noe sikkerhetssignal i denne utprøvingen og forbedret dokumentasjonen. Høy tillit til konklusjonen om at det ikke påviselig forbedret pasientutfallene etter 14 dager her. Lav tillit til enhver påstand om at det «virker» eller «mislykkes» som et tiltak for pasientnytte — det spørsmålet er genuint uavklart og krever en langt større studie. En passende holdning: Et grundig resultat fra den virkelige verden om et verktøy som ikke ga noe sikkerhetssignal og forbedret behandlingsprosessen, ikke en dom om at KI forvandler — eller ødelegger — primærhelsetjenesten.

Kilder

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