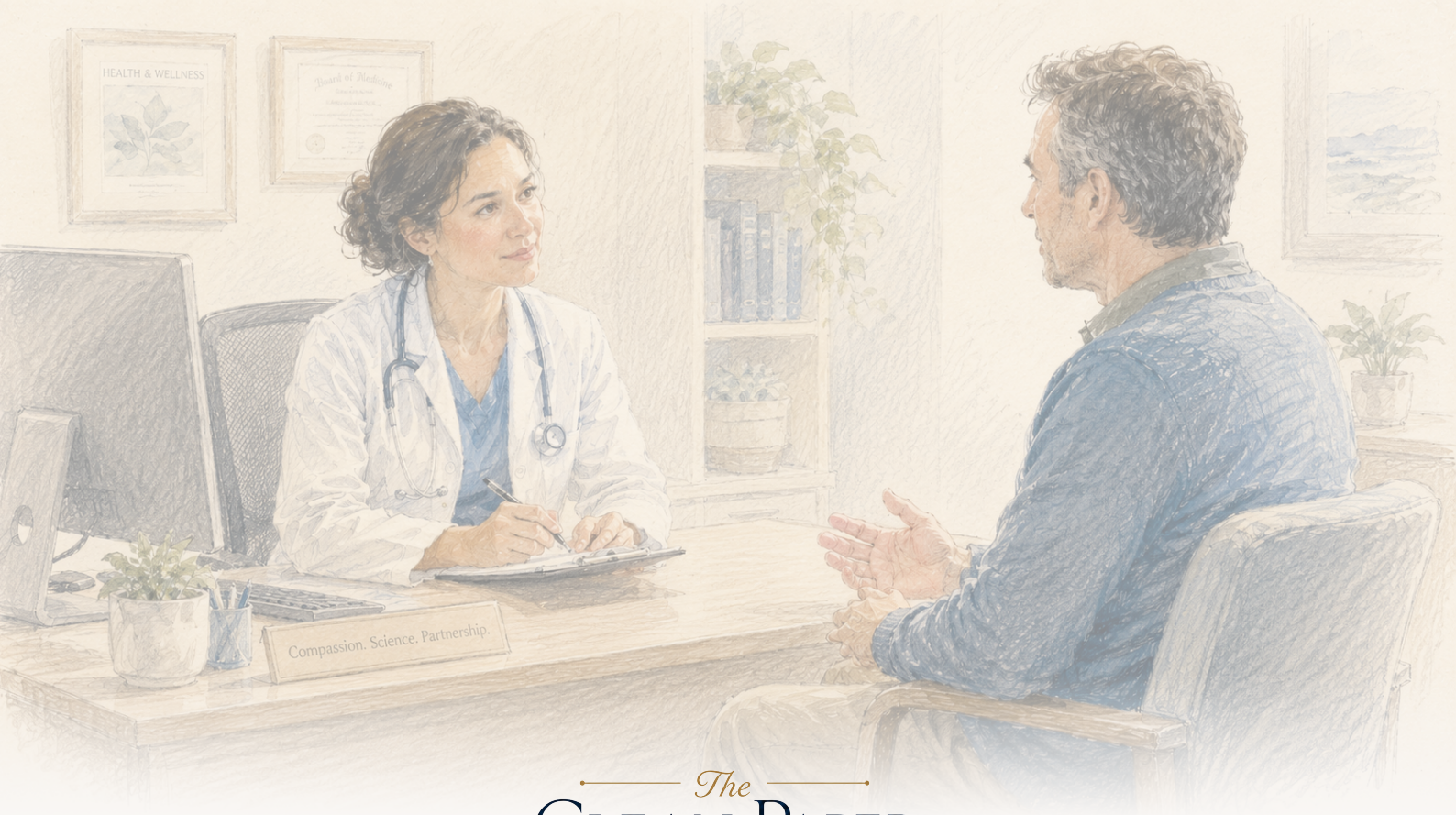


The Clean Paper

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

En klinisk AI, der så sikker ud og forbedrede papirarbejdet — men ikke patientresultaterne

Et pragmatisk, klyngerandomiseret forsøg på 16 kenyanske primærklinikker testede et beslutningsstøtteværktøj med generativ AI, som blev føjet til den journal, clinical officers brugte. Blandt 9.691 patienter var det strenge, ekspertbedømte, sammensatte mål for behandlingssvigt efter 14 dage 2,2 % med værktøjet mod 2,0 % uden (justeret oddsratio 0,77, 95 % KI 0,55-1,08, P = 0,13) — ingen signifikant forskel. Værktøjet viste intet sikkerhedssignal, forbedrede dokumentationen på alle bedømte områder, lod ordinationer og tilfredshed være uændrede og viste en tendens til lavere antibiotikaomkostninger. Det er ikke et mislykket studie; det er et sjældent, ærligt studie — målt på patienter, ikke på benchmarktest — og det lander på den uglamourøse sandhed, at et værktøj, som så sikkert ud og hjalp processen, ikke påviseligt hjalp patienterne på to uger, og at det ville kræve et langt større forsøg at opdage en eventuel sådan gevinst.

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

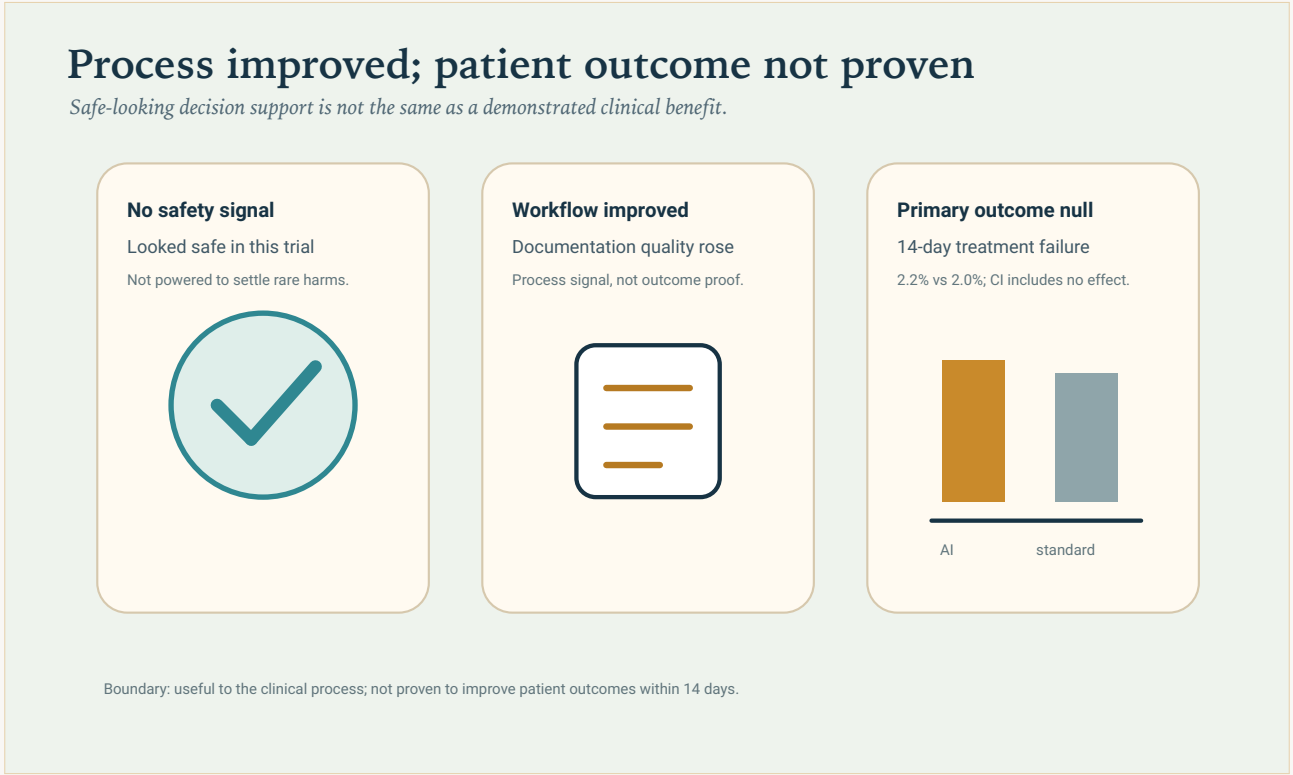
Sikkert og nyttigt er ikke det samme som effektivt

De fleste overskrifter om medicinsk AI bygger på den forkerte slags studie. En model klarer sig godt på eksamensspørgsmål eller slår læger på kuraterede patientvignetter, og historien skriver sig selv: Maskinen er klar til klinikken.

Dette forsøg gjorde noget sværere og sjældnere. Det satte et beslutningsstøtteværktøj med generativ AI ind i den virkelige primærsektor, på virkelige klinikker, med virkelige patienter, og stillede det spørgsmål, der faktisk betyder noget: Fik patienterne det bedre?

Det ærlige svar er nej — ikke målbart, ikke inden for 14 dage. Værktøjet viste intet sikkerhedssignal. Det forbedrede kvaliteten af den kliniske dokumentation. Det kan endda have sænket visse lægemiddelomkostninger. Men det reducerede ikke behandlingssvigt signifikant, og forfatterne er omhyggelige med at sige, at en eventuel gevinst sandsynligvis er beskeden.

Det er ikke et mislykket studie. Det er studiet, der virker. Sådan ser ansvarlig evidens om AI i medicin ud, når den måles på patienter i stedet for på benchmarktest.



Værktøjet viste intet sikkerhedssignal og forbedrede den kliniske dokumentation, men det på forhånd specificerede patientresultat efter 14 dage blev ikke signifikant bedre. Hjælp i processen er ikke det samme som dokumenteret patientgevinst.

Hvad forfatterne gjorde

Forskerholdet gennemførte et pragmatisk, klyngerandomiseret forsøg på **16 primærklinikker**, der drives af et privat sundhedsnetværk (Penda Health) i amterne Nairobi og Kiambu i Kenya. Behandlingen på disse klinikker varetages i vid udstrækning af **clinical officers** — sundhedspersonale på mellemniveau med en treårig diplomuddannelse — ofte uden let adgang til at konsultere en mere erfaren kollega.

Randomiseringsenheden var klinikeren, ikke patienten. **103 clinical officers** blev randomiseret: 52 til interventionsarmen og 51 til kontrolarmen. Begge arme brugte den samme cloudbaserede elektroniske patientjournal. Interventionsarmen havde desuden **»AI Consult«** (version 2.0), et beslutningsstøtteværktøj bygget på den store sprogmodel **OpenAI's GPT-4o** og integreret i journalen. Det læste de oplysninger, klinikeren dokumenterede, og kunne gøre opmærksom på mulige problemer med diagnosen eller behandlingsplanen. Klinikerne bevarede fuld selvbestemmelse: De kunne acceptere, ændre eller ignorere forslagene.

> **Hvilken model var det, og hvorfor detaljerne betyder noget**
>
> Værktøjet var **AI Consult 2.0**, som kørte **OpenAI's GPT-4o** (udgaven fra maj 2025), blev tilgået via OpenAI's kommercielle API med en virksomhedslicens og kørte med indstillinger for lav tilfældighed (temperatur 0,1). Det var integreret i en skræddersyet elektronisk journal (EasyClinic's EMR) og styret af systeminstrukser skrevet, så de stemte overens med Kenyas nationale behandlingsretningslinjer; forfatterne offentliggjorde hele instruktionsprompten.
> Hvorfor præcisere alt dette? Fordi resultatet handler om *ét bestemt system* — én modelversion, én prompt, én journal, én indstilling — ikke om »store sprogmodeller i medicin« generelt. Forfatterne fremhæver det samme: De kalder deres fund et *tidsbestemt benchmark snarere end et fast estimat af kapaciteten*. En nyere model, en anden prompt eller en mindre digitaliseret klinik kunne alle ændre resultatet.
> Om uafhængighed: OpenAI ydede senere naturalier (kreditter til cloudcomputing og teknisk vejledning i brugen af virksomhedens API), men forfatterne oplyser, at beslutningen om at bruge OpenAI blev truffet *før* dette tilbud, og at OpenAI ikke spillede nogen rolle i forsøgets design, dataindsamling, analyse eller beslutningen om at offentliggøre.

Mellem 22. april og 16. juli 2025 blev **9.691 patienter** inkluderet. Det **primære resultat var bevidst patientcentreret og strengt**: et ekspertbedømt *sammensat mål for behandlingssvigt* inden for 14 dage efter konsultationen — et panel af klinikere, som var blindet for studiearm, vurderede, om hver patient havde haft et dårligt resultat, eksempelvis en sygdom, der ikke var forsvundet eller var blevet forværret. Forsøget var registreret på forhånd (Pan-African Clinical Trials Registry 202502499779176).

Dette designvalg er selve pointen. Det er let at vise, at et AI-værktøj ændrer, hvad en kliniker skriver ned. Det er langt sværere, og langt mere meningsfuldt, at vise, at det ændrer, hvad der sker med patienten.

Hvad de fandt

Det primære resultat blev ikke bedre. Behandlingssvigt forekom hos 102 af 4.693 patienter (2,2 %) i AI-armen og 94 af 4.654 (2,0 %) i kontrolarmen. De ujusterede procenttal var marginalt højere med AI, men efter justering pegede punktestimatet i retning af en gevinst: Den **justerede oddsratio var 0,77 (95 % konfidensinterval 0,55 til 1,08, P = 0,13)** — ikke statistisk signifikant. At retningen skifter mellem de rå og justerede tal, er ikke en regnefejl; justeringen tager højde for forskelle mellem klyngerne af klinikere. Uanset hvad indeholder konfidensintervallet med god margin »ingen effekt«, så der kan ikke hævdes nogen gevinst, og i absolutte tal var effekten meget lille.

For en lettilgængelig forklaring på, hvordan oddsratioer, konfidensintervaller og P-værdier skal læses sammen, se vejledningen til kliniske resultater.

Intet sikkerhedssignal — inden for visse grænser. Ingen alvorlige bivirkninger blev vurderet til at være relateret til værktøjet, og en uafhængig gennemgang fandt intet sikkerhedssignal. Forfatterne er ærlige om grænsen for denne tryghed: Forsøget havde ikke statistisk styrke til at opdage sjældne, alvorlige skader, og det havde ingen på forhånd specificeret ramme for noninferioritet eller formel sikkerhed. Det kan derfor ikke *bevise* sikkerhed ved usædvanlige hændelser.

Dokumentationen blev bedre. Blandt 2.000 konsultationer, som blev gennemgået af blindede eksperter, udarbejdede klinikere, der brugte AI Consult, bedre klinisk dokumentation på alle bedømte områder — den registrerede diagnose, behandlingsplanen og den samlede fuldstændighed.

Ordinationerne ændrede sig næsten ikke. Der var ingen signifikant forskel i ordinationer, heller ikke i korrekt brug af antibiotika (justeret oddsratio 0,86, 95 % KI 0,48 til 1,55). Værktøjet ændrede ikke andelen af antibiotikaordinationer.

Patienterne bemærkede ingen forskel. Blandt 826 patienter, som gennemførte en tilfredshedsundersøgelse, var tilfredsheden stort set identisk mellem armene, og konsultationstiderne var omtrent de samme.

Omkostningerne pegede svagt nedad. I en justeret analyse var de antibiotikarelaterede omkostninger lavere i AI-armen — sandsynligvis gennem billigere valg snarere end færre ordinationer — og besparelsen på antibiotika pr. patient så ud til at overstige omkostningen pr. patient ved at drive værktøjet. Forfatterne betegner dette som et fingerpeg, ikke som afgjort: En fuldstændig opgørelse af de samlede ejeromkostninger lå uden for forsøget.

Forfatternes egen sammenfatning i én sætning er den klareste version: LLM-støtten var sikker inden for disse grænser, men reducerede ikke behandlingssvigt inden for 14 dage, og en eventuel gevinst er sandsynligvis beskeden.

Hvorfor »ingen signifikant forskel« ikke betyder »det virker ikke«

Et nulresultat for det primære resultat er let at overfortolke i begge retninger. To forhold står i vejen for den enkle fortælling.

For det første var **forsøget tilrettelagt til at opfange en større effekt end den, det fandt.** Alvorlige dårlige resultater i primærsektoren er sjældne — omkring 2 % her — så det kræver enorme deltagerantal at skelne en lille, reel gevinst fra støj. Forfatternes egne post hoc-beregninger af statistisk styrke tyder på, at det ville kræve et **langt større forsøg, i størrelsesordenen mere end 100.000 patienter**, at opdage en effekt af den størrelse, de observerede. Et ikke-signifikant resultat i en undersøgelse af denne størrelse udelukker ikke en lille, reel gevinst; det betyder, at denne undersøgelse ikke kunne afgøre, om der var en.

For det andet var **sammenligningen delvist udvisket.** En konfigurationsfejl gav kortvarigt nogle klinikere i kontrolarmen adgang til AI Consult, og klinikere i et fælles netværk taler med hinanden og tager vaner med sig på tværs af grænsen. Begge dele har en tendens til at gøre de to arme mere ens og dermed skubbe en eventuel reel forskel mod nul. Dertil kommer, at værtsnetværket allerede arbejdede efter relativt høje standarder, hvilket giver et værktøj mindre plads til at vise forbedring.

Intet af dette redder en overskrift om et »gennembrud«. Men det betyder, at den korrekte læsning er afstemt, ikke afvisende: På det vanskeligste og mest ærlige endepunkt kunne det ikke påvises, at dette værktøj hjalp patienter på to uger — samtidig med at det målbart hjalp journalføringen og så sikkert ud.

Hvad dette *ikke* beviser

- Det viser **ikke**, at AI'en forbedrede patientresultaterne. På det primære endepunkt efter 14 dage var der ingen signifikant gevinst.
- Det viser **ikke**, at AI'en er ubrugelig. Punktestimatet talte til dens fordel, dokumentationen blev bedre over hele linjen, og lægemiddelomkostningerne havde en faldende tendens; nulresultatet er foreneligt med en lille, reel gevinst, som forsøget var for lille til at bekræfte.
- Det beviser **ikke**, at værktøjet er sikkert med hensyn til sjældne skader. Det viste intet sikkerhedssignal, men forsøget havde hverken statistisk styrke eller et design, der kunne dokumentere sikkerhed ved usædvanlige, alvorlige hændelser.
- Det viser **ikke**, at »AI slår læger« eller erstatter klinikere. Dette er beslutnings*støtte*; klinikeren bevarede fuld myndighed til at acceptere eller afvise den.
- Det kan **ikke** automatisk generaliseres. Forsøget blev gennemført i ét privat bynetværk i Kenya; forhold i landdistrikter, bynære områder og højindkomstmiljøer kan afvige i begge retninger.
- Det fastslår **ikke** omkostningsbesparelser. Omkostningssignalet er et fingerpeg, ikke en fuldført sundhedsøkonomisk evaluering.

Hvor stærk er evidensen?

For den centrale påstand — ingen dokumenteret reduktion i kortsigtet patientskade — er evidensen stærk *som design* og passende beskeden *som konklusion*. Et prospektivt, forhåndsregistreret, klyngerandomiseret forsøg med et blindet, sammensat resultat på patientniveau er tæt på den bedste evidens fra den virkelige verden, man kan indsamle for et værktøj som dette. Det er langt mere informativt end en score på en benchmarktest eller en undersøgelse med patientvignetter.

For de sekundære fund — bedre dokumentation, uændrede ordinationer, samme tilfredshed, lavere antibiotikaomkostninger — er evidensen god, men bør læses som sekundær: understøttende signaler, ikke hovedbudskabet, og sårbare over for de samme begrænsninger med kontaminering og et enkelt netværk.

For sikkerhed er evidensen betryggende, men afgrænset: Der blev ikke fundet noget signal, men dette var ikke en undersøgelse tilrettelagt til at finde sjældne skader.

Den mest nyttige holdning er hverken »det virker« eller »det mislykkedes«. Den er: Et omhyggeligt forsøg i den virkelige verden fandt, at dette AI-værktøj ikke gav noget sikkerhedssignal og forbedrede behandlingsprocessen uden at påvise nogen gevinst for patientresultaterne på to uger — og at det ville kræve en langt større undersøgelse at opdage en eventuel sådan gevinst.

Hvorfor det er vigtigt

Debatten om medicinsk AI mangler netop denne slags evidens. Der findes tusindvis af artikler, som viser modeller, der klarer eksamener til topkarakter og matcher klinikere i velordnede cases. Der findes meget få store, pragmatiske, randomiserede forsøg, som måler, om virkelige patienter får det bedre. Dette er et af dem, og det lander på den uglamourøse sandhed: At bestå prøven er ikke det samme som at hjælpe patienten.

Det mellemrum er hele historien. Et værktøj kan være oprigtigt nyttigt for klinikere — tydeligere notater, lavere lægemiddelregninger, et ekstra sæt øjne — og alligevel ikke flytte et krævende patientresultat på to uger. Begge dele kan være sandt samtidig, og et modent sundhedsvæsen skal holde dem sammen i stedet for at vælge det bekvemme.

Det flytter også bevisbyrden i en nyttig retning. Hvis en virksomhed vil hævde, at dens kliniske AI forbedrer behandlingen, er den relevante evidens ikke en rangliste. Det er et forsøg som dette med resultater, der betyder noget for patienter — og helst et større et, fordi den ærlige lære her er, at beskedne gevinster kræver store undersøgelser for at kunne opdages.

Kort sammenfatning

Et pragmatisk, klyngerandomiseret forsøg på 16 kenyanske primærklinikker testede et beslutningsstøtteværktøj med generativ AI (»AI Consult«), som blev føjet til den elektroniske journal, clinical officers brugte. Blandt 9.691 patienter var det ekspertbedømte, sammensatte mål for behandlingssvigt inden for 14 dage 2,2 % med værktøjet mod 2,0 % uden (justeret oddsratio 0,77, 95 % KI 0,55–1,08, P = 0,13) — ingen signifikant forskel. Værktøjet viste intet sikkerhedssignal, forbedrede den kliniske dokumentation på alle bedømte områder, ændrede ikke ordinationerne, lod patienttilfredsheden være uændret og var forbundet med noget lavere antibiotikaomkostninger. Forfatterne konkluderer, at det var sikkert inden for disse grænser, men ikke reducerede behandlingssvigt, og at en eventuel gevinst sandsynligvis er beskeden. Det ville kræve et langt større forsøg (i størrelsesordenen 100.000 patienter) at opdage en effekt af den observerede størrelse. Resultatet viser hverken, at klinisk AI forbedrer patientresultater, eller at den er ubrugelig — det viser, at et værktøj, der så sikkert ud og hjalp processen, ikke påviseligt hjalp patienterne på to uger i ét privat bynetværk.

No-BS-tjek

Hvad artiklen viser: I et randomiseret forsøg i den virkelige verden gav tilføjelsen af et LLM-baseret beslutningsstøtteværktøj til patientjournalerne i primærsektoren intet sikkerhedssignal, forbedrede kvaliteten af den kliniske dokumentation, ændrede ikke ordinationerne og reducerede ikke signifikant et strengt, sammensat mål for behandlingssvigt hos patienter efter 14 dage.

Hvad der er plausibelt, men ikke bevist: At værktøjet giver en lille, reel reduktion i behandlingssvigt, som var for lille til, at dette forsøg kunne opdage den; at det sparer penge, når alle omkostninger medregnes; at bedre dokumentation med tiden fører til bedre behandling.

Hvad den ikke viser: At klinisk AI forbedrer patientresultater; at den er usikker ved sjældne hændelser; at den erstatter eller overgår klinikere; at disse resultater kan overføres til landdistrikter eller højindkomstmiljøer; at omkostningsbesparelsen er fastslået.

Vigtigste begrænsninger: Dimensioneret til en større effekt end den observerede (sjældne resultater kræver meget store stikprøver); en konfigurationsfejl kontaminerede kontrolarmen, og vaner i det fælles netværk udvisker sammenligningen, hvilket i begge tilfælde forskyder resultatet mod ingen forskel; ét privat bynetværk med allerede høje standarder; ingen på forhånd specificeret ramme for noninferioritet eller sikkerhed; kort tidshorisont på 14 dage.

Hvor stor tillid bør en almindelig læser have? Stor tillid til, at værktøjet ikke viste noget sikkerhedssignal i dette forsøg og forbedrede dokumentationen. Stor tillid til konklusionen om, at det ikke påviseligt forbedrede patientresultaterne efter 14 dage her. Lille tillid til enhver påstand om, at det »virker« eller »mislykkes« som en intervention til patientgevinst — det spørgsmål er reelt uafklaret og kræver en langt større undersøgelse. En passende holdning: Et omhyggeligt resultat fra den virkelige verden om et værktøj, som ikke gav noget sikkerhedssignal og forbedrede behandlingsprocessen, ikke en dom om, at AI forvandler — eller ødelægger — primærsektoren.

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