The Clean Paper

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

Een klinische AI die veilig leek en de administratie verbeterde — maar de patiëntuitkomsten niet

Een pragmatische, cluster-gerandomiseerde trial in 16 Keniaanse eerstelijnslocaties testte een generatieve-AI-beslisondersteuner in het dossier van clinical officers. Onder 9.691 patiënten was het strenge, door experts beoordeelde samengestelde eindpunt voor behandelingsfalen binnen 14 dagen 2,2% met het hulpmiddel versus 2,0% zonder (gecorrigeerde oddsratio 0,77, 95%-BI 0,55–1,08, P = 0,13) — geen significant verschil. Het hulpmiddel liet geen veiligheidssignaal zien, verbeterde documentatie in alle beoordeelde domeinen, liet voorschrijven en tevredenheid onveranderd en wees op iets lagere antibioticakosten. Dat is geen mislukte studie maar een zeldzaam eerlijke: gemeten op patiënten, niet op benchmarks.

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

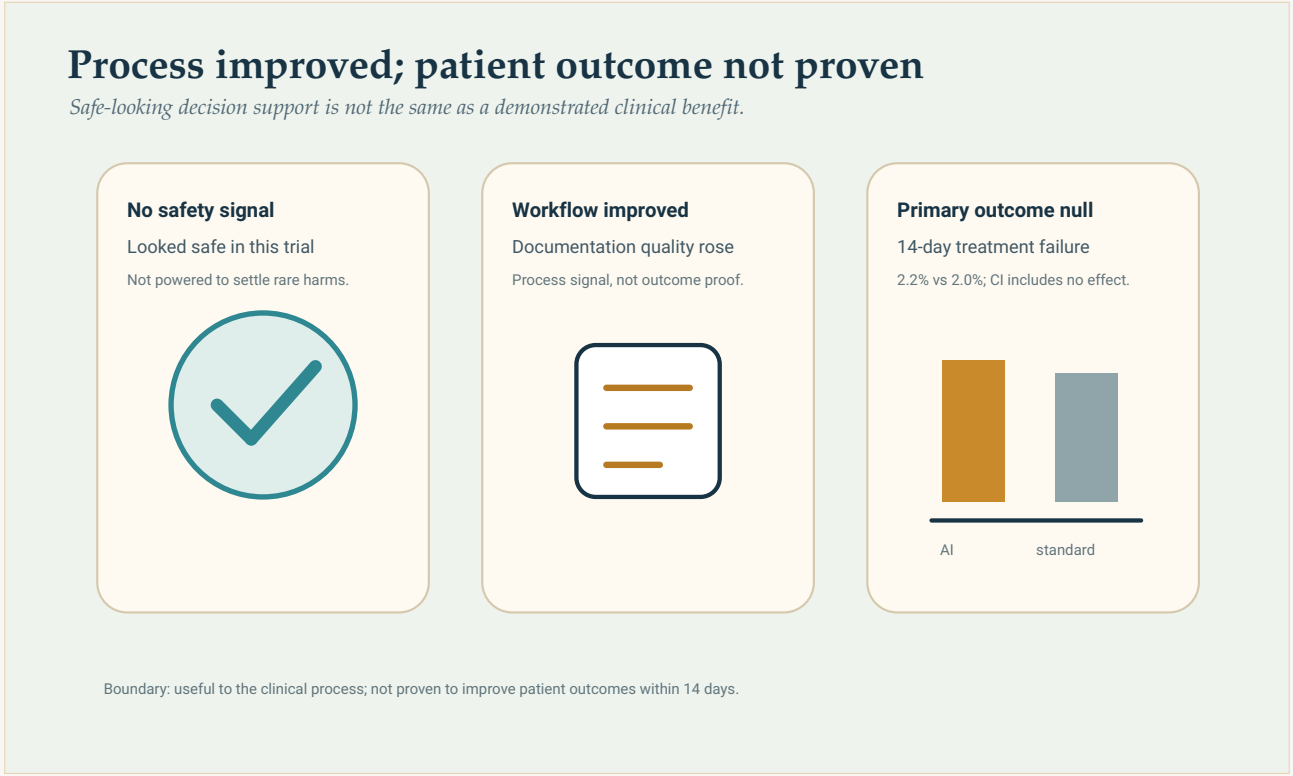
Veilig en behulpzaam is niet hetzelfde als effectief

De meeste krantenkoppen over medische AI zijn gebaseerd op het verkeerde soort onderzoek. Een model scoort goed op examenvragen of verslaat artsen op zorgvuldig geselecteerde casussen, en het verhaal schrijft zichzelf: de machine is klaar voor de kliniek.

Deze trial deed iets moeilijkers en zeldzamers. Een generatieve-AI-beslisondersteuner werd in echte eerstelijnszorg gezet, in echte klinieken, bij echte patiënten, en vervolgens werd de vraag gesteld die er werkelijk toe doet: ging het beter met de patiënten?

Het eerlijke antwoord is nee — niet meetbaar, niet binnen 14 dagen. Het hulpmiddel liet geen veiligheidssignaal zien. Het verbeterde de kwaliteit van de klinische documentatie. Mogelijk verlaagde het zelfs sommige medicijnkosten. Maar het verminderde het aantal behandelingsmislukkingen niet significant, en de auteurs zeggen voorzichtig dat een eventueel voordeel waarschijnlijk bescheiden is.

Dat is geen mislukking van de studie. Dat is de studie die haar werk doet. Zo ziet verantwoord bewijs over AI in de geneeskunde eruit wanneer je op patiënten meet in plaats van op benchmarks.



Het hulpmiddel liet geen veiligheidssignaal zien en verbeterde de klinische documentatie, maar het vooraf vastgelegde patiënteneindpunt na 14 dagen verbeterde niet significant. Proceswinst is niet hetzelfde als bewezen patiëntvoordeel.

Wat de auteurs deden

Het team voerde een pragmatische, cluster-gerandomiseerde trial uit in **16 eerstelijnszorglocaties** van het particuliere zorgnetwerk Penda Health in de Keniaanse counties Nairobi en Kiambu. De zorg wordt daar grotendeels geleverd door **clinical officers** — zorgprofessionals op middelbaar niveau met een driejarige opleiding — vaak zonder gemakkelijke toegang tot een senior voor overleg.

De randomisatie-eenheid was de behandelaar, niet de patiënt. **103 clinical officers** werden gerandomiseerd: 52 naar de interventiegroep en 51 naar de controlegroep. Beide groepen gebruikten hetzelfde cloudgebaseerde elektronische patiëntendossier. De interventiegroep kreeg daarnaast **AI Consult** (versie 2.0), een beslisondersteunend hulpmiddel op basis van **OpenAI's GPT-4o**, geïntegreerd in dat dossier. Het las wat de behandelaar documenteerde en kon mogelijke problemen met diagnose of behandelplan signaleren. De behandelaar behield volledige autonomie en kon suggesties accepteren, aanpassen of negeren.

> **Welk model was het, en waarom zijn de details belangrijk**
>
> Het hulpmiddel was **AI Consult 2.0**, draaiend op **OpenAI's GPT-4o** (release van mei 2025), via OpenAI's commerciële API onder een enterprise-licentie en ingesteld op lage willekeur (temperature 0,1). Het zat in een op maat gemaakt elektronisch dossier (EasyClinic's EMR) en werd gestuurd door systeemprompts die waren geschreven om aan te sluiten bij de Keniaanse nationale behandelrichtlijnen; de auteurs publiceerden de volledige instructieprompt.
> Waarom dit zo precies benoemen? Omdat het resultaat over *één specifiek systeem* gaat — één modelversie, één prompt, één dossier, één instelling — niet over "LLM's in de geneeskunde" in het algemeen. De auteurs maken hetzelfde punt: ze noemen hun bevinding een *tijdgebonden benchmark in plaats van een vaste schatting van capaciteit*. Een nieuwer model, andere prompt of minder gedigitaliseerde kliniek kan een andere uitkomst geven.
> Wat onafhankelijkheid betreft: OpenAI leverde later ondersteuning in natura (cloudcomputecredits en technische begeleiding bij het gebruik van de API), maar volgens de auteurs was de keuze voor OpenAI al gemaakt *vóór* dat aanbod en had OpenAI geen rol in het ontwerp, de dataverzameling, de analyse of de beslissing om te publiceren.

Tussen 22 april en 16 juli 2025 werden **9.691 patiënten** geïncludeerd. Het **primaire eindpunt was bewust patiëntgericht en streng**: een door experts beoordeelde *samengestelde uitkomst van behandelingsfalen* binnen 14 dagen na het bezoek. Een panel clinici, geblindeerd voor de studiearm, beoordeelde of een patiënt een slechte uitkomst had, zoals aanhoudende of verergerde ziekte. De trial was vooraf geregistreerd (Pan-African Clinical Trials Registry 202502499779176).

Die ontwerpkeuze is precies het punt. Het is gemakkelijk om te laten zien dat een AI-hulpmiddel verandert wat een behandelaar opschrijft. Het is veel moeilijker — en veel betekenisvoller — om te laten zien dat het verandert wat er met de patiënt gebeurt.

Wat ze vonden

Het primaire eindpunt verbeterde niet. Behandelingsfalen trad op bij 102 van 4.693 patiënten (2,2%) in de AI-groep en 94 van 4.654 (2,0%) in de controlegroep. De ruwe percentages waren dus iets hoger met AI, maar na correctie schoof de puntschatting juist richting voordeel: de **gecorrigeerde oddsratio was 0,77 (95%-BI 0,55 tot 1,08, P = 0,13)** — niet statistisch significant. Dat de richting tussen ruwe en gecorrigeerde cijfers omdraait is geen rekenfout; de correctie houdt rekening met verschillen tussen de clusters van behandelaars. Hoe dan ook omvat het betrouwbaarheidsinterval ruim "geen effect", dus er kan geen voordeel worden geclaimd, en in absolute zin was het effect klein.

Voor een gewone uitleg van oddsratio's, betrouwbaarheidsintervallen en P-waarden samen, zie de gids voor het lezen van klinische resultaten.

Geen veiligheidssignaal — binnen grenzen. Geen ernstig ongewenst voorval werd aan het hulpmiddel toegeschreven en een onafhankelijke beoordeling vond geen veiligheidssignaal. De auteurs zijn eerlijk over de grens van die geruststelling: de trial was niet groot genoeg om zeldzame ernstige schade op te sporen en had geen vooraf vastgelegde non-inferioriteits- of formele veiligheidsopzet. Ze kan dus niet *bewijzen* dat zeldzame ernstige gebeurtenissen geen probleem zijn.

De documentatie werd beter. Van 2.000 consulten die door geblindeerde experts werden beoordeeld, leverden behandelaars met AI Consult betere klinische documentatie in alle beoordeelde domeinen: diagnose, behandelplan en algemene volledigheid.

Het voorschrijfgedrag veranderde nauwelijks. Er was geen significant verschil in voorschrijven, inclusief correct antibioticagebruik (gecorrigeerde oddsratio 0,86, 95%-BI 0,48 tot 1,55). Het hulpmiddel veranderde de antibioticavoorschrijffrequentie niet.

Patiënten merkten geen verschil. Onder 826 patiënten die een tevredenheidsenquête invulden, was de tevredenheid vrijwel gelijk in beide groepen en waren de consulttijden vergelijkbaar.

De kosten wezen licht naar beneden. In een gecorrigeerde analyse waren antibioticagerelateerde kosten lager in de AI-groep — plausibel door goedkopere keuzes en niet door minder voorschriften — en de besparing per patiënt leek groter dan de kosten per patiënt voor het draaien van het hulpmiddel. De auteurs behandelen dit als suggestief, niet als bewezen: een volledige total-cost-of-ownershipanalyse viel buiten de trial.

De eigen samenvatting van de auteurs is de schoonste: LLM-ondersteuning was binnen deze grenzen veilig, maar verminderde behandelingsfalen binnen 14 dagen niet, en enig voordeel is waarschijnlijk bescheiden.

Waarom “geen significant verschil” niet hetzelfde is als “het werkt niet”

Een nulresultaat op het primaire eindpunt is makkelijk in beide richtingen te overinterpreteren. Twee dingen blokkeren het simpele verhaal.

Ten eerste was **de trial ontworpen om een groter effect te vinden dan er werd gezien**. Ernstige slechte uitkomsten in de eerste lijn zijn zeldzaam — hier ongeveer 2% — en een klein werkelijk voordeel van ruis onderscheiden vereist enorme aantallen. De eigen post-hoc powerberekeningen van de auteurs suggereren dat voor een effect van de waargenomen grootte een **veel grotere trial nodig zou zijn, in de orde van meer dan 100.000 patiënten**. Een niet-significant resultaat in deze studie sluit dus een klein echt voordeel niet uit; het betekent dat deze studie zo’n voordeel niet kon oplossen.

Ten tweede was **de vergelijking deels vervaagd**. Door een configuratiefout hadden sommige behandelaars in de controlegroep tijdelijk toegang tot AI Consult. Bovendien praten behandelaars binnen hetzelfde netwerk met elkaar en nemen ze gewoonten mee over de groepsgrens. Beide effecten maken de groepen meer op elkaar gelijk en duwen een werkelijk verschil richting nul. Daarbovenop draaide het netwerk al op relatief hoge kwaliteitsnormen, waardoor minder ruimte overblijft voor verbetering.

Geen van deze punten redt een “doorbraak”-kop. Maar ze betekenen wel dat de juiste lezing gekalibreerd moet zijn, niet defaitistisch: op het moeilijkste en eerlijkste eindpunt hielp dit hulpmiddel patiënten binnen twee weken niet aantoonbaar — terwijl het de documentatie wel meetbaar verbeterde en er veilig uitzag.

Wat dit *niet* bewijst

- Het laat **niet** zien dat AI de patiëntuitkomsten verbeterde. Op het primaire eindpunt na 14 dagen was er geen significant voordeel.
- Het laat **niet** zien dat AI nutteloos is. De puntschatting ging richting voordeel, documentatie verbeterde overal en medicijnkosten gingen licht omlaag; het nulresultaat is verenigbaar met een klein werkelijk voordeel dat de trial niet kon bevestigen.
- Het **bewijst geen veiligheid voor zeldzame schade**. Er was geen veiligheidssignaal, maar de studie was niet ontworpen of groot genoeg om zeldzame ernstige gebeurtenissen uit te sluiten.
- Het laat **niet** zien dat “AI artsen verslaat” of clinici vervangt. Dit is *beslisondersteuning*; de behandelaar behield de volledige bevoegdheid om het advies te volgen of te negeren.
- Het **generaliseert niet automatisch**. De trial liep in één particulier stedelijk netwerk in Kenia; rurale, peri-urbane en rijkere zorgsystemen kunnen anders uitpakken.
- Het **bewijst geen kostenbesparing**. Het kostensignaal is suggestief, geen volledige economische evaluatie.

Hoe sterk is het bewijs?

Voor de centrale claim — geen bewezen vermindering van kortetermijnschade voor patiënten — is het bewijs sterk *qua ontwerp* en terecht bescheiden *qua conclusie*. Een prospectieve, vooraf geregistreerde, cluster-gerandomiseerde trial met een geblindeerd samengesteld eindpunt op patiëntniveau zit dicht bij het beste real-world bewijs dat je voor zo'n hulpmiddel kunt verzamelen. Het is veel informatiever dan een benchmarkscore of een studie met vignetten.

Voor de secundaire bevindingen — betere documentatie, onveranderd voorschrijven, vergelijkbare tevredenheid, lagere antibioticakosten — is het bewijs goed, maar het blijven secundaire signalen: ondersteunend, niet de hoofdclaim, en kwetsbaar voor dezelfde contaminatie en beperkingen van één netwerk.

Voor veiligheid is het bewijs geruststellend maar begrensd: geen signaal gevonden, maar geen studie die zeldzame schade moest vinden.

De nuttigste houding is dus noch "het werkt" noch "het faalde". Het is: een zorgvuldige trial in de echte wereld vond geen veiligheidssignaal en zag betere zorgprocessen, maar toonde binnen twee weken geen voordeel voor patiëntuitkomsten — en om zo'n mogelijk voordeel te zien is een veel grotere studie nodig.

Waarom het ertoe doet

Het debat over medische AI heeft juist aan dit soort bewijs gebrek. Er zijn duizenden papers waarin modellen examens halen en op nette casussen naast clinici worden gezet. Er zijn maar weinig grote, pragmatische, gerandomiseerde trials die meten of echte patiënten er beter van worden. Dit is er één, en de uitkomst is de onglamoureuze waarheid: slagen voor de test is niet hetzelfde als de patiënt helpen.

Die kloof is het hele verhaal. Een hulpmiddel kan werkelijk nuttig zijn voor clinici — duidelijkere notities, lagere medicijnrekening, een tweede paar ogen — en toch binnen veertien dagen geen hard patiënteneindpunt verplaatsen. Beide kunnen tegelijk waar zijn. Een volwassen zorgsysteem moet ze samen kunnen vasthouden in plaats van alleen het handige feit te kiezen.

Het zet ook de bewijslast nuttig terug op zijn plaats. Als een bedrijf beweert dat zijn klinische AI de zorg verbetert, is het relevante bewijs geen leaderboard. Het is een trial als deze, met uitkomsten die patiënten aangaan — en idealiter een grotere, omdat de eerlijke les hier is dat bescheiden voordelen enorme studies vragen.

Heldere samenvatting

Een pragmatische, cluster-gerandomiseerde trial in 16 Keniaanse eerstelijnszorglocaties testte een generatieve-AI-beslisondersteuner (“AI Consult”) die werd toegevoegd aan het elektronische dossier van clinical officers. Onder 9.691 patiënten was het door experts beoordeelde samengestelde eindpunt behandelingsfalen binnen 14 dagen 2,2% met het hulpmiddel en 2,0% zonder (gecorrigeerde oddsratio 0,77, 95%-BI 0,55–1,08, P = 0,13) — geen significant verschil. Het hulpmiddel liet geen veiligheidssignaal zien, verbeterde klinische documentatie in alle beoordeelde domeinen, veranderde het voorschrijfgedrag niet, liet patiënttevredenheid onveranderd en hing samen met iets lagere antibioticakosten. De auteurs concluderen dat het binnen die grenzen veilig was maar behandelingsfalen niet verminderde, waarbij enig voordeel waarschijnlijk bescheiden is; een effect van de waargenomen grootte zou een veel grotere trial vergen, in de orde van 100.000 patiënten. Het resultaat laat niet zien dat klinische AI patiëntuitkomsten verbetert, maar evenmin dat het nutteloos is — wel dat een hulpmiddel dat veilig leek en het proces verbeterde patiënten in twee weken niet aantoonbaar hielp, in één particulier stedelijk netwerk.

Zonder omwegen

Wat het artikel laat zien: In een gerandomiseerde real-world trial veroorzaakte toevoeging van een LLM-beslisondersteuner aan eerstelijnsdossiers geen veiligheidssignaal, verbeterde zij de kwaliteit van klinische documentatie, veranderde zij voorschrijven niet en verminderde zij een streng samengesteld patiënteneindpunt voor behandelingsfalen binnen 14 dagen niet significant.

Wat aannemelijk is maar niet bewezen: Dat het hulpmiddel een klein werkelijk voordeel geeft dat deze trial niet kon detecteren; dat het geld bespaart wanneer alle kosten worden meegeteld; dat betere documentatie later in betere zorg resulteert.

Wat het niet laat zien: Dat klinische AI patiëntuitkomsten verbetert; dat het onveilig is voor zeldzame gebeurtenissen; dat het clinici vervangt of overtreft; dat deze resultaten vanzelf gelden in rurale of rijkere settings; dat de kostenbesparing vaststaat.

Belangrijkste beperkingen: De trial was berekend op een groter effect dan waargenomen (zeldzame uitkomsten vragen zeer grote steekproeven); een configuratiefout besmette de controlegroep en gedeelde gewoonten binnen het netwerk vervagen de vergelijking, beide richting geen verschil; één particulier stedelijk netwerk met al hoge standaarden; geen vooraf vastgelegde non-inferioriteits- of veiligheidsopzet; korte horizon van 14 dagen.

Hoeveel vertrouwen mag een algemene lezer hierin hebben? Hoog vertrouwen dat het hulpmiddel in deze trial geen veiligheidssignaal liet zien en documentatie verbeterde. Hoog vertrouwen dat het hier geen aantoonbare verbetering van patiëntuitkomsten na 14 dagen gaf. Laag vertrouwen in elke algemene conclusie dat het als patiëntinterventie “werkt” of “faalt” — die vraag is werkelijk onopgelost en vereist een veel grotere studie. De gepaste houding: een zorgvuldige real-world uitkomst over een hulpmiddel dat veilig leek en het zorgproces verbeterde, geen oordeel dat AI de eerste lijn transformeert — of vernielt.

Bronnen

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