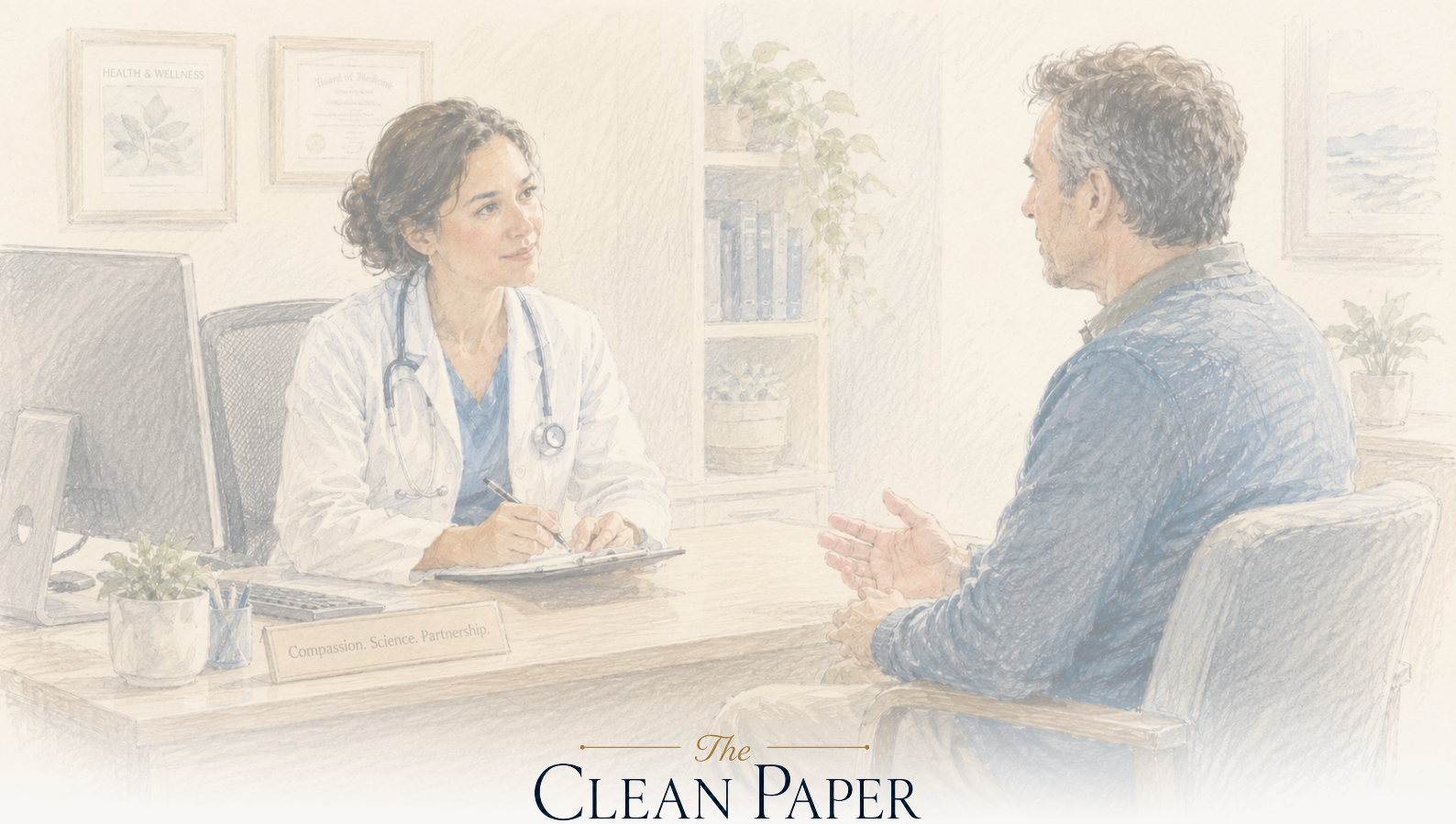


The Clean Paper

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

'n Kliniese KI wat veilig gelyk en die papierwerk verbeter het — maar nie pasiëntuitkomste nie

'n Pragmatiese, tros-gerandomiseerde proef in 16 Keniaanse primêresorgfasiliteite het 'n generatiewe-KI-besluitsteuninstrument getoets wat by die rekord van kliniese beamptes gevoeg is. Onder 9 691 pasiënte was die streng, kundig beoordeelde 14-dae-saamgestelde maatstaf van behandelingsmislukking 2,2% met die instrument teenoor 2,0% daarsonder (aangepaste kansverhouding 0,77, 95%-VI 0,55–1,08, P = 0,13) — geen beduidende verskil nie. Die instrument het geen veiligheidssein getoon nie, dokumentasie in al die beoordeelde domeine verbeter, voorskryfgedrag en tevredenheid onveranderd gelaat en na laer antibiotikakoste geneig. Dit is nie 'n mislukte studie nie; dit is 'n seldsame, eerlike een — op pasiënte gemeet, nie op maatstawwe nie — en dit beland by die onglansryke waarheid dat 'n instrument wat veilig gelyk en die proses gehelp het, pasiënte nie aantoonbaar binne twee weke gehelp het nie, en dat enige sulke voordeel 'n veel groter proef sou verg om op te spoor.

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

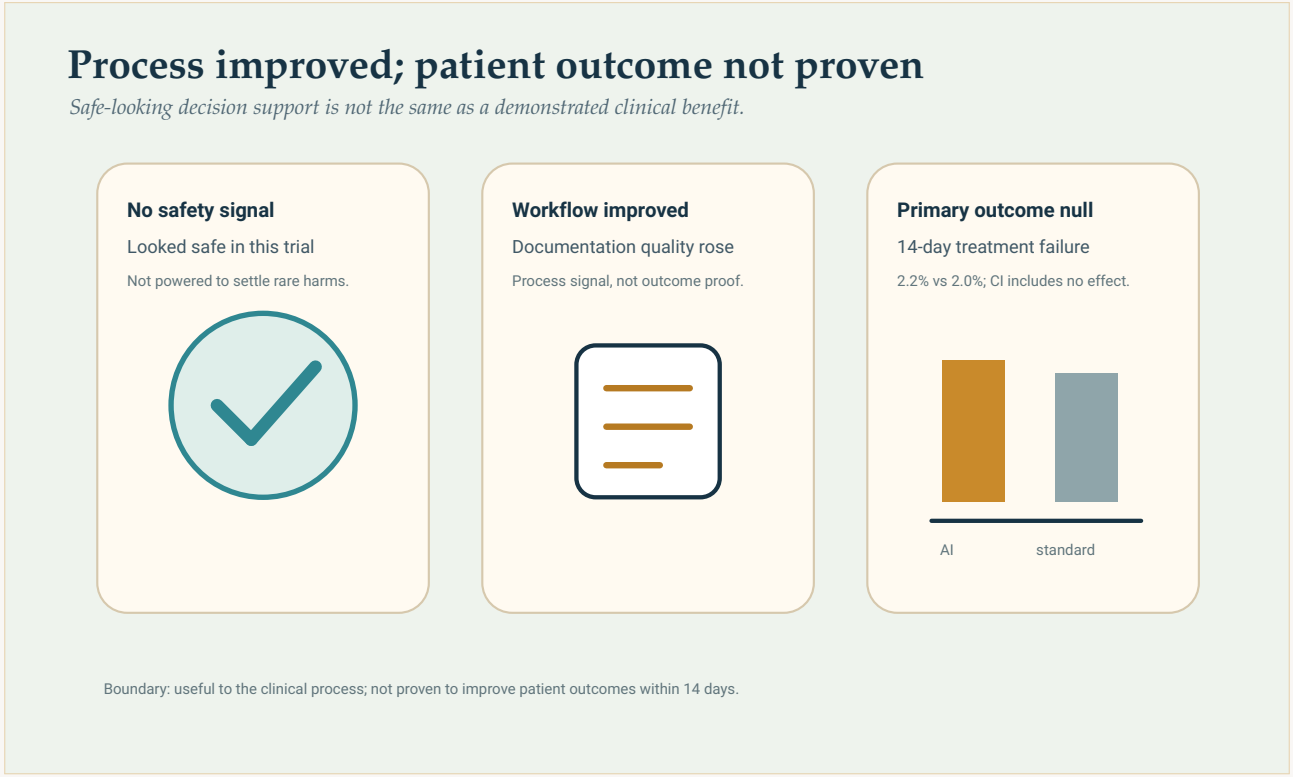
Veilig en nuttig is nie dieselfde as doeltreffend nie

Die meeste opskrifte oor mediese KI is op die verkeerde soort studie gebou. ’n Model vaar goed met eksamenvrae, of klop dokters op netjies saamgestelde gevalle, en die verhaal skryf homself: die masjien is gereed vir die kliniek.

Hierdie proef het iets moeiliker en seldsamer gedoen. Dit het ’n generatiewe-KI-besluitsteuninstrument in werklike primêre sorg geplaas, in werklike fasiliteite, met werklike pasiënte, en die vraag gevra wat werklik saak maak: het pasiënte beter gevaar?

Die eerlike antwoord is nee — nie meetbaar nie, nie binne 14 dae nie. Die instrument het geen veiligheidssein getoon nie. Dit het die gehalte van kliniese dokumentasie verbeter. Dit kon selfs sommige medisynekoste verlaag het. Maar dit het behandelingsmislukkings nie beduidend verminder nie, en die outeurs is versigtig om te sê dat enige voordeel, as dit bestaan, waarskynlik beskeie is.

Dit is nie ’n mislukking van die studie nie. Dit is die studie wat werk. Só lyk verantwoordelike bewys oor KI in medisyne wanneer dit op pasiënte eerder as op maatstawwe gemeet word.



Die instrument het geen veiligheidssein getoon nie en kliniese dokumentasie verbeter, maar die voorafbepaalde 14-dae-pasiëntuitkoms het nie beduidend verbeter nie. Hulp met die proses is nie dieselfde as bewese voordeel vir pasiënte nie.

The Clean Paper CC BY 4.0

Wat die outeurs gedoen het

Die span het ’n pragmatiese, tros-gerandomiseerde proef oor **16 primêresorgfasiliteite** van ’n private gesondheidsnetwerk (Penda Health) in Nairobi- en Kiambu-graafskappe, Kenia, uitgevoer. Sorg in dié fasiliteite word grootliks deur **kliniese beamptes** gelewer — middelvlak-praktisyns met ’n driejaar-diploma — dikwels sonder maklike toegang tot senior konsultasie.

Die eenheid van randomisering was die klinikus, nie die pasiënt nie. **103 kliniese beamptes** is gerandomiseer: 52 na die intervensie-arm en 51 na die kontrole-arm. Albei arms het dieselfde wolkgebaseerde elektroniese mediese rekord gebruik. Die intervensie-arm het daarby **“AI Consult”** (weergawe 2.0) gehad, ’n besluitsteuninstrument wat op **OpenAI se GPT-4o**-groottaalmodel gebou en in dié rekord ingebed is. Dit het die inligting gelees wat ’n klinikus aangeteken het en kon moontlike probleme met die diagnose of behandelingsplan uitwys. Klinici het volle outonomie behou: hulle kon die voorstelle aanvaar, verander of ignoreer.

> **Watter model was dit, en waarom die besonderhede saak maak**
>
> Die instrument was **AI Consult 2.0**, wat **OpenAI se GPT-4o** (die Mei 2025-uitgawe) gebruik het, via OpenAI se kommersiële API onder ’n ondernemingslisensie en met laewillekeurinstellings (temperatuur 0,1). Dit was binne ’n pasgemaakte elektroniese rekord (EasyClinic se EMR) ingebed en is met stelselopdragte gelei wat op Kenia se nasionale behandelingsriglyne afgestem is; die outeurs het die volledige instruksie gepubliseer.
> Waarom dit uitspel? Omdat die resultaat oor *een spesifieke stelsel* gaan — een modelweergawe, een opdrag, een rekordstelsel, een instelling — nie oor “LLM’s in medisyne” in die algemeen nie. Die outeurs maak dieselfde punt: hulle noem hul bevinding ’n *tydgebonde maatstaf eerder as ’n vaste skatting van vermoë*. ’n Nuwer model, ’n ander opdrag of ’n minder gedigitaliseerde kliniek kan almal die uitkoms verander.
> Wat onafhanklikheid betref: OpenAI het later ondersteuning in natura verskaf (krediete vir wolkrekenaargebruik en tegniese leiding oor die API), maar die outeurs sê die besluit om OpenAI te gebruik is *voor* dié aanbod geneem, en dat OpenAI geen rol in die proefontwerp, dataversameling, ontleding of besluit om te publiseer gehad het nie.

Tussen 22 April en 16 Julie 2025 is **9 691 pasiënte** ingeskryf. Die **primêre uitkoms was doelbewus pasiëntgesentreerd en streng**: ’n deur kundiges beoordeelde *saamgestelde maatstaf van behandelingsmislukking* binne 14 dae ná die besoek — ’n paneel klinici het, blind vir die studie-arm, beoordeel of elke pasiënt ’n slegte uitkoms gehad het, soos siekte wat nie opgelos het nie of vererger het. Die proef is vooraf geregistreer (Pan-African Clinical Trials Registry 202502499779176).

Daardie ontwerpkeuse is die punt. Dit is maklik om te wys dat ’n KI-instrument verander wat ’n klinikus neerskryf. Dit is baie moeiliker, en baie betekenisvoller, om te wys dat dit verander wat met die pasiënt gebeur.

Wat hulle gevind het

Die primêre uitkoms het nie verbeter nie. Behandelingsmislukking het voorgekom by 102 van 4 693 pasiënte (2,2%) in die KI-arm en 94 van 4 654 (2,0%) in die kontrole-arm. Die ru persentasies was fraksioneel hoër met KI, maar ná aanpassing het die puntskatting in die rigting van voordeel geleun: die **aangepaste kansverhouding was 0,77 (95%-vertrouensinterval 0,55 tot 1,08, P = 0,13)** — nie statisties beduidend nie. Dat die ru en aangepaste syfers in teenoorgestelde rigtings wys, is nie ’n rekenfout nie; die aanpassing neem verskille tussen die klinikustrosse in ag. Hoe dit ook al sy, die vertrouensinterval sluit “geen effek” gemaklik in, dus kan geen voordeel geëis word nie, en in absolute terme was die effek klein.

Vir ’n gewone-taalgids oor hoe om kansverhoudings, vertrouensintervalle en P-waardes saam te lees, sien die gids tot kliniese resultate.

Geen veiligheidssein nie — binne perke. Geen ernstige nadelige gebeurtenisse is as verwant aan die instrument beoordeel nie, en ’n onafhanklike oorsig het geen veiligheidssein gevind nie. Die outeurs is eerlik oor die beperking op dié gerusstelling: die proef was nie groot genoeg om seldsame ernstige skade op te spoor nie en het geen voorafbepaalde nie-minderwaardigheids- of formele veiligheid-sraamwerk gehad nie; dit kan dus nie veiligheid vir ongewone ernstige gebeurtenisse *bewys* nie.

Die dokumentasie het beter geword. Onder 2 000 konsultasies wat deur geblindeerde kundiges nagegaan is, het klinici wat AI Consult gebruik het beter kliniese dokumentasie gelewer in al die beoordeelde domeine — die aangetekende diagnose, die behandelingsplan en algehele volledigheid.

Voorskryfgedrag het skaars verander. Daar was geen beduidende verskil in voorskryfgedrag nie, insluitend korrekte antibiotikagebruik (aangepaste kansverhouding 0,86, 95%-VI 0,48 tot 1,55). Die instrument het nie die tempo van antibiotikavoorskryf verander nie.

Pasiënte het geen verskil opgemerk nie. Onder 826 pasiënte wat ’n tevredenheidsopname voltooi het, was tevredenheid feitlik identies tussen die arms, en konsultasietye was soortgelyk.

Koste het effens afwaarts gewys. In ’n aangepaste ontleding was antibiotikaverwante koste laer in die KI-arm — waarskynlik deur goedkoper keuses eerder as minder voorskrifte — en die besparing per pasiënt op antibiotika het groter gelyk as die koste per pasiënt om die instrument te bedryf. Die outeurs merk dit as suggestief, nie afgehandel nie: ’n volledige rekening van totale eienaarskoste het buite die proef geval.

Die outeurs se eie eenreël-opsomming is die skoonste weergawe: LLM-bystand was veilig binne dié perke maar het behandelingsmislukking binne 14 dae nie verminder nie, en enige voordeel is waarskynlik beskeie.

Waarom “geen beduidende verskil” nie “dit werk nie” beteken nie

’n Nul primêre uitkoms is maklik om in albei rigtings te oorlees. Twee dinge keer die eenvoudige verhaal.

Eerstens: **die proef is ontwerp om ’n groter effek te vind as dié wat waargeneem is.** Ernstige slegte uitkomste in primêre sorg is skaars — hier ongeveer 2% — dus verg dit enorme getalle om ’n klein werklike voordeel van geraas te onderskei. Die outeurs se eie post-hoc-kragberekeninge dui daarop dat ’n effek van die waargenome grootte ’n **veel groter proef, in die orde van meer as 100 000 pasiënte**, sou vereis. ’n Nie-beduidende resultaat in ’n studie van hierdie grootte sluit nie ’n klein werklike voordeel uit nie; dit beteken dat hierdie studie dit nie kon uitklaar nie.

Tweedens: **die vergelyking is gedeeltelik vertroebel.** ’n Konfigurasiefout het sommige klinici in die kontrole-arm kortliks toegang tot AI Consult gegee, en klinici in ’n gedeelde netwerk praat met mekaar en dra gewoontes oor die grens heen. Albei effekte laat die twee arms meer na mekaar lyk en druk enige werklike verskil in die rigting van nul. Daarby het die gasheernetwerk reeds relatief hoë standaarde gehandhaaf, wat minder ruimte vir verbetering laat.

Niks hiervan red ’n “deurbraak”-opskrif nie. Maar dit beteken wel dat die korrekte lees gekalibreer, nie afbrekend, is: op die moeilikste en eerlikste eindpunt het hierdie instrument pasiënte nie aantoonbaar binne twee weke gehelp nie — terwyl dit die rekordhouding wel meetbaar verbeter het en veilig gelyk het.

Wat dit *nie* bewys nie

- Dit wys **nie dat die KI pasiëntuitkomste verbeter het nie**. Op die primêre 14-dae-eindpunt was daar geen beduidende voordeel nie.
- Dit wys **nie dat die KI nutteloos is nie**. Die puntskatting het dit bevoordeel, dokumentasie het oor die hele linie verbeter en medisynekoste het afwaarts geneig; die nulresultaat is verenigbaar met ’n klein werklike voordeel wat die proef te klein was om te bevestig.
- Dit **bewys nie dat die instrument veilig is vir seldsame skade nie**. Dit het geen veiligheidssein getoon nie, maar die studie was nie groot genoeg of ontwerp om veiligheid vir ongewone ernstige gebeurtenisse te sertifiseer nie.
- Dit wys **nie dat “KI dokters klop” of klinici vervang nie**. Dit is besluit*steun*; die klinikus het volle gesag behou om dit te aanvaar of te verwerp.
- Dit **veralgemeen nie outomaties nie**. Die proef het in een private stedelike netwerk in Kenia plaasgevind; landelike, peri-stedelike en hoërinkomste-omgewings kan in enige rigting verskil.
- Dit **vestig nie kostebesparing nie**. Die kostesein is suggestief, nie ’n voltooide ekonomiese evaluering nie.

Hoe sterk is die bewys?

Vir die sentrale aanspraak — geen bewese afname in korttermyn-pasiëntskade nie — is die bewys sterk *as ontwerp*, en toepaslik beskeie *as gevolgtrekking*. ’n Prospektiewe, voorafgeregistreerde, trosgerandomiseerde proef met ’n geblindeerde, pasiëntvlak-saamgestelde uitkoms is naby aan die beste werklike bewys wat vir so ’n instrument versamel kan word. Dit is veel meer insiggewend as ’n maatstaftelling of ’n studie van netjiese gevalbeskrywings.

Vir die sekondêre bevindings — beter dokumentasie, onveranderde voorskryfgedrag, soortgelyke tevredenheid, laer antibiotikakoste — is die bewys goed, maar behoort dit as sekondêr gelees te word: ondersteunende seine, nie die hoofopskrif nie, en onderhewig aan dieselfde kontaminasie- en enkelnetwerkbeperkings.

Vir veiligheid is die bewys gerusstellend maar begrens: geen sein is gevind nie, maar die studie was nie gebou om seldsame skade te vind nie.

Die nuttigste houding is nóg “dit werk” nóg “dit het misluk”. Dit is: ’n versigtige werklike proef het gevind dat hierdie KI-instrument geen veiligheidssein opgewek het nie en die sorgproses verbeter het, sonder om ’n voordeel vir pasiëntuitkomste binne twee weke te demonstreer — en om so ’n voordeel te vind sou ’n veel groter studie verg.

Waarom dit saak maak

Die debat oor mediese KI het presies hierdie soort bewys bitter nodig. Daar is duisende artikels wat wys hoe modelle eksamens uitmuntend slaag en klinici op netjiese gevalle ewenaar. Daar is baie min groot, pragmatiese, gerandomiseerde proewe wat meet of werklike pasiënte beter daaraan toe is. Hierdie is een daarvan, en dit beland by die onglansryke waarheid: om die toets te slaag is nie dieselfde as om die pasiënt te help nie.

Daardie gaping is die hele verhaal. ’n Instrument kan werklik nuttig vir klinici wees — duideliker notas, laer medisynekoste, ’n tweede paar oë — en steeds nie binne twee weke ’n harde pasiëntuitkoms skuif nie. Albei feite kan tegelyk waar wees, en ’n volwasse gesondheidstelsel moet hulle saam kan hou in plaas daarvan om net die gerieflike een te kies.

Dit stel ook die bewyslas op ’n nuttige manier terug. As ’n maatskappy wil beweer dat sy kliniese KI sorg verbeter, is die relevante bewys nie ’n ranglys nie. Dit is ’n proef soos hierdie, op uitkomste wat vir pasiënte saak maak — en verkieslik ’n groter een, want die eerlike les hier is dat beskeie voordele groot studies verg om raak te sien.

Kortliks

'n Pragmatiese, tros-gerandomiseerde proef in 16 Keniaanse primêresorgfasiliteite het 'n generatiewe-KI-besluitsteuninstrument ("AI Consult") getoets wat by die elektroniese rekord van kliniese beamptes gevoeg is. Onder 9 691 pasiënte was die deur kundiges beoordeelde saamgestelde maatstaf van behandelingsmislukking binne 14 dae 2,2% met die instrument teenoor 2,0% daarsonder (aangepaste kansverhouding 0,77, 95%-VI 0,55–1,08, P = 0,13) — geen beduidende verskil nie. Die instrument het geen veiligheidssein getoon nie, kliniese dokumentasie in al die beoordeelde domeine verbeter, voorskryfgedrag nie verander nie, pasiënttevredenheid onveranderd gelaat en was met ietwat laer antibiotikakoste geassosieer. Die outeurs kom tot die gevolgtrekking dat dit binne dié perke veilig was maar behandelingsmislukking nie verminder het nie, met enige voordeel waarskynlik beskeie; om 'n effek van die waargenome grootte op te spoor sou 'n veel groter proef verg, in die orde van 100 000 pasiënte. Die resultaat wys nie dat kliniese KI pasiëntuitkomste verbeter nie, en ook nie dat dit nutteloos is nie — dit wys dat 'n instrument wat veilig gelyk en die proses gehelp het, pasiënte nie aantoonbaar binne twee weke gehelp het in een private stedelike netwerk nie.

Nugtere beoordeling

Wat die artikel wys: In 'n werklike gerandomiseerde proef het die toevoeging van 'n LLM-gebaseerde besluitsteuninstrument tot primêresorgrekords geen veiligheidssein opgewek nie, die gehalte van kliniese dokumentasie verbeter, voorskryfgedrag nie verander nie en 'n streng 14-dae-saamgestelde maatstaf van behandelingsmislukking by pasiënte nie beduidend verminder nie.

Wat aannemelik is maar nie bewys nie: Dat die instrument 'n klein werklike afname in behandelingsmislukking lewer wat te klein was vir hierdie proef om op te spoor; dat dit geld bespaar wanneer alle koste ingereken word; dat beter dokumentasie uiteindelik in beter sorg vertaal.

Wat dit nie wys nie: Dat kliniese KI pasiëntuitkomste verbeter; dat dit veilig is vir seldsame gebeurtenisse; dat dit klinici vervang of beter as hulle presteer; dat dié resultate na landelike of hoërinkomste-omgewings oordra; dat die kostebesparing vasgestel is.

Belangrikste beperkings: Die proef was bemagtig vir 'n groter effek as dié wat waargeneem is (skaars uitkomste verg baie groot monsters); 'n konfigurasiefout het die kontrole-arm gekontamineer en gewoontes binne 'n gedeelde netwerk vertroebel die vergelyking, wat albei die resultaat na geen verskil toe druk; een private stedelike netwerk met reeds hoë standaarde; geen voorafbepaalde nie-minderwaardigheids- of veiligheidsraamwerk nie; kort horison van 14 dae.

Hoeveel vertroue behoort 'n algemene leser hierin te hê? Hoë vertroue dat die instrument in hierdie proef geen veiligheidssein getoon en dokumentasie verbeter het nie. Hoë vertroue dat dit pasiëntuitkomste binne 14 dae hier nie aantoonbaar verbeter het nie. Lae vertroue in enige aanspraak dat dit as 'n pasiëntvoordeel-intervensie "werk" of "misluk" — daardie vraag is werklik onbeslis en verg 'n veel groter studie. Gepaste houding: 'n versigtige werklike resultaat oor 'n instrument wat geen veiligheidssein opgewek het nie en die sorgproses verbeter het, nie 'n uitspraak dat KI primêre sorg

transformeer — of verwoes — nie.

Bronne

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