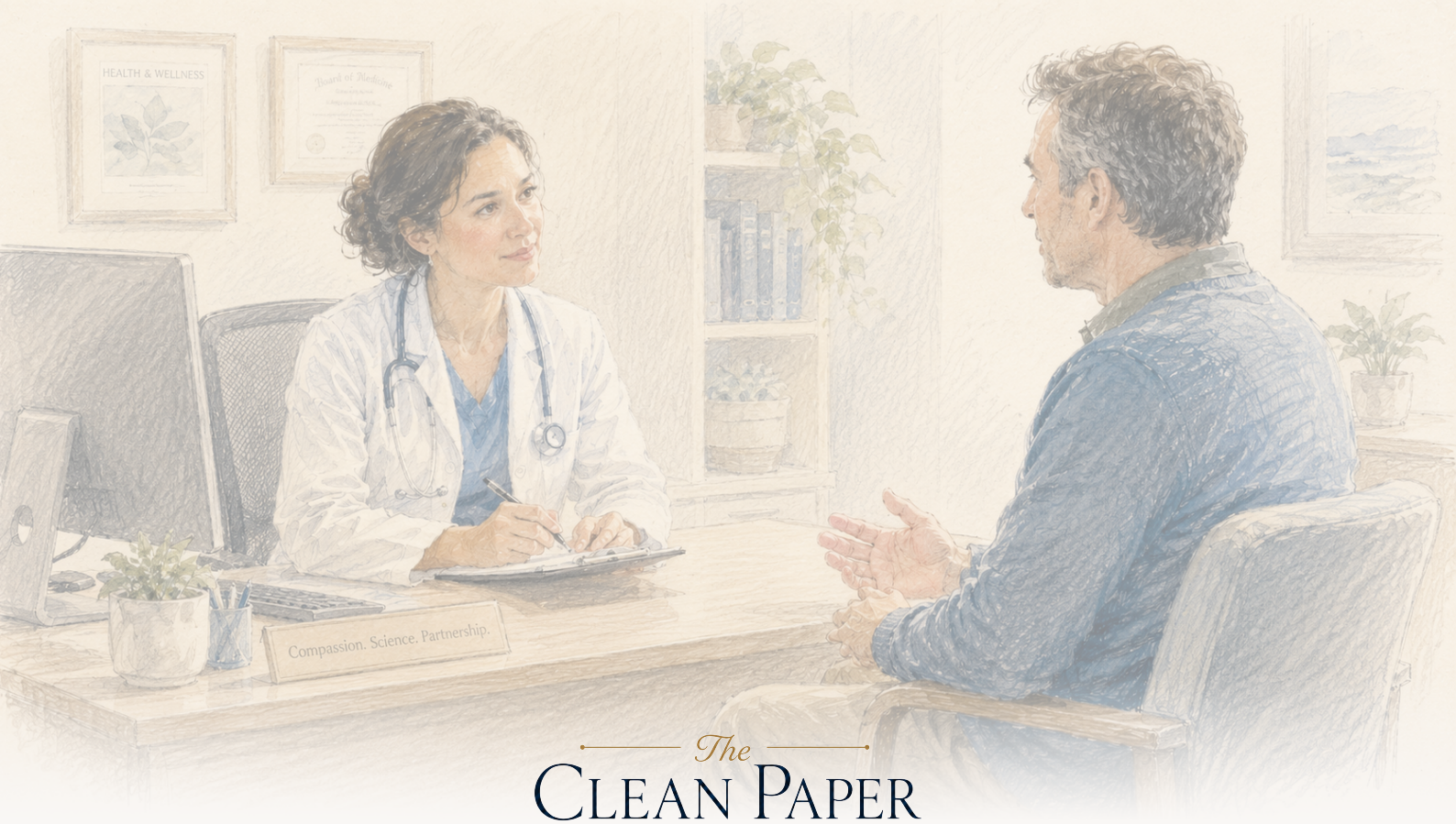


The Clean Paper

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

Klinički AI koji je izgledao sigurno i poboljšao dokumentaciju — ali nije poboljšao ishode pacijenata

Pragmatično, klasterski randomizovano ispitivanje u 16 kenijskih ustanova primarne zdravstvene zaštite testiralo je alat za podršku odlučivanju zasnovan na generativnoj veštačkoj inteligenciji, dodan kartonu kojim su se koristili klinički službenici. Među 9,691 pacijenata, strogi 14-dnevni složeni ishod neuspeha lečenja koji su procenjivali stručnjaci iznosio je 2.2% uz alat i 2.0% bez njega (prilagođeni odnos izgleda 0.77, 95% CI 0.55-1.08, P = 0.13) — bez značajne razlike. Alat nije pokazao bezbednosni signal, poboljšao je dokumentaciju u svim ocenjivanim oblastima, nije promenio propisivanje ni zadovoljstvo i pokazao je trend prema nižim troškovima antibiotika. To nije neuspela studija; to je retka, poštena studija — merena na pacijentima, a ne na benčmarkovima — koja završava na neglamuroznoj istini da alat koji je izgledao sigurno i pomagao procesu nije dokazivo pomogao pacijentima za dve nedelje te da bi za otkrivanje eventualne koristi trebalo mnogo veće ispitivanje.

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

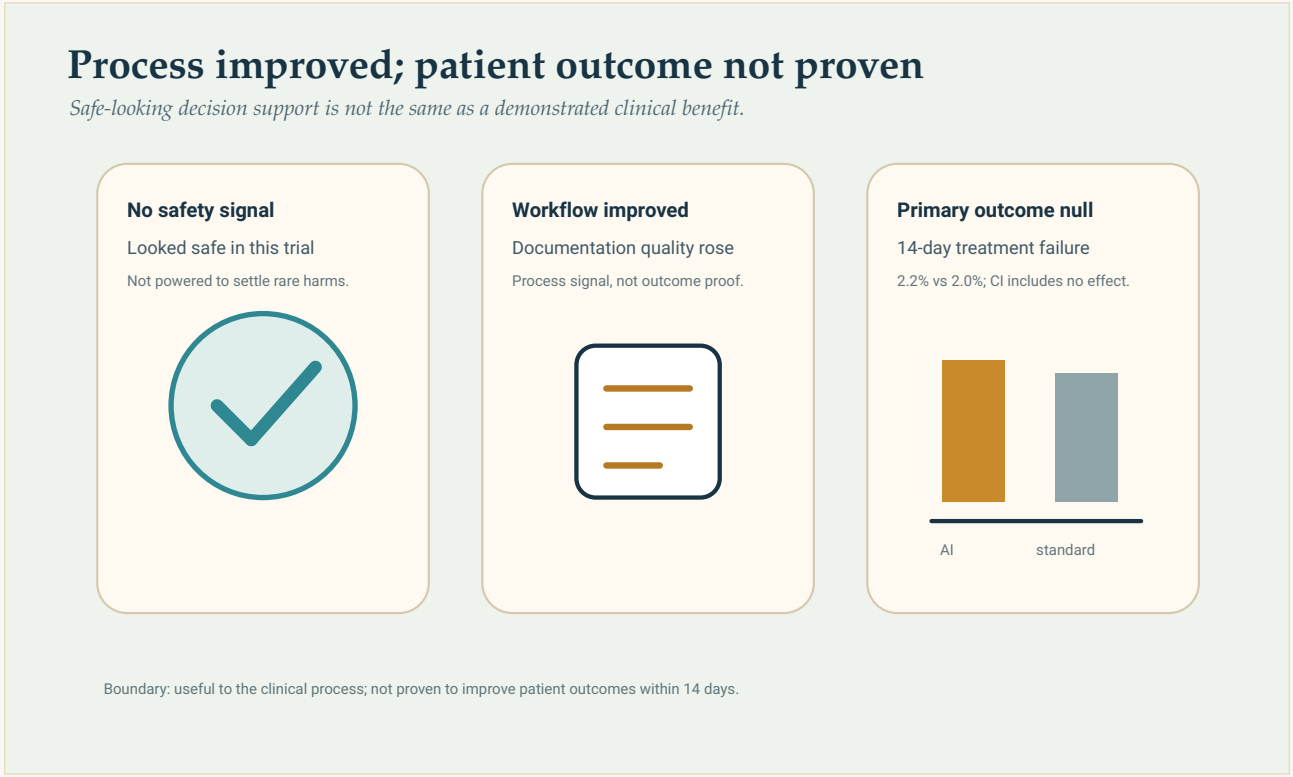
Sigurno i korisno nije isto što i efikasno

Većina naslova o medicinskoj veštačkoj inteligenciji zasniva se na pogrešnoj vrsti studije. Model dobro reši ispitna pitanja ili nadmaši lečnike na pažljivo odabranim kliničkim vinjetama i priča se napiše sama: mašina je spremna za ambulantu.

Ovo ispitivanje napravilo je nešto teže i ređe. Uvelo je alat za podršku odlučivanju zasnovan na generativnoj veštačkoj inteligenciji u stvarnu primarnu zdravstvenu zaštitu, u stvarne ustanove, sa stvarnim pacijentima, i postavilo pitanje koje je zaista važno: da li su pacijenti prošli bolje?

Pošten odgovor je ne — barem ne merljivo i ne unutar 14 dana. Alat nije pokazao bezbednosni signal. Poboljšao je kvalitet kliničke dokumentacije. Možda je čak smanjio deo troškova lekova. Ali nije statistički značajno smanjio neuspehe lečenja, a autori oprezno navode da je eventualna korist verovatno skromna.

To nije neuspeh studije. To je studija koja radi svoj posao. Ovako izgleda odgovorno prikupljanje dokaza o veštačkoj inteligenciji u medicini kada se meri na pacijentima, a ne na benčmarkovima.



Alat nije pokazao bezbednosni signal i poboljšao je kliničku dokumentaciju, ali unapred određeni ishod za pacijente nakon 14 dana nije se statistički značajno poboljšao. Pomoć procesu nije isto što i dokazana korist za pacijenta.

Šta su autori uradili

Tim je sproveo pragmatično, klasterski randomizovano ispitivanje u **16 ustanova primarne zdravstvene zaštite** privatne zdravstvene mreže Penda Health u okruzima Nairobi i Kiambu u Keniji. Nega u tim ustanovama uglavnom pružaju **klinički službenici** — zdravstveni radnici srednjeg nivoa s trogodišnjom stručnim obrazovanjem — često bez lakog pristupa konsultacijama sa starijim kolegama.

Jedinica randomizacije bio je kliničar, a ne pacijent. Randomizovana su **103 klinička službenika**: 52 u intervencijsku i 51 u kontrolnu grupu. Obe grupe koristile su isti elektronski zdravstveni karton u oblaku. Intervencijska grupa dodatno je imala **AI Consult** (verzija 2.0), alat za podršku odlučivanju izgrađen na velikom jezičnom modelu **OpenAI GPT-4o** i ugrađen u taj karton. Čitao je podatke koje je kliničar unio i mogao upozoriti na moguće probleme s dijagnozom ili planom lečenja. Kliničari su zadržali potpunu autonomiju: mogli su prihvatiti, izmeniti ili zanemariti predloge.

> **Koji je model bio u pitanju i zašto su pojedinosti važne**
>
> Alat je bio **AI Consult 2.0**, koji je koristio **OpenAI GPT-4o** (izdanje iz maja 2025.) preko OpenAI-jeva komercijalnog API-ja pod poslovnom licencom i pri postavkama male nasumičnosti (temperature 0.1). Bio je ugrađen u prilagođeni elektronski karton (EasyClinicov EMR) i vođen sistemskim promptovima napisanima tako da slede kenijske nacionalne smernice lečenja; autori su objavili celi instrukcioni prompt.
> zašto ovo toliko precizirati? Zato što se rezultat odnosi na *jedan određeni sistem* — jednu verziju modela, jedan prompt, jedan karton i jednu konfiguraciju — a ne na „LLM-ove u medicini" uopšteno. Autori naglašavaju isto: svoj rezultat nazivaju *vremenskim benčmarkom, a ne fiksnom procenom sposobnosti*. Noviji model, drugačiji prompt ili manje digitalizovana ambulanta mogli bi promeniti ishod.
> Što se tiče nezavisnosti: OpenAI je naknadno pružio nenovčanu podršku (kredite za računarske resurse u oblaku i tehničke savete za korišćenje API-ja), ali autori navode da je odluka o korišćenju OpenAI-ja donesena *pre* te ponude te da OpenAI nije imao ulogu u dizajnu ispitivanja, prikupljanju podataka, analizi ni odluci o objavi.

Između 22. aprila i 16. jula 2025. uključeno je **9,691 pacijenata**. **Primarni ishod namerno je bio strogo usmeren na pacijente**: stručno procenjen *složeni ishod neuspeha lečenja* unutar 14 dana od posete — panel kliničara, zaslepljen za grupu kojoj je pacijent pripadao, procenjivao je da li je pacijent imao loš ishod poput nerešene ili pogoršane bolesti. Ispitivanje je unapred registrovano (Pan-African Clinical Trials Registry 202502499779176).

Upravo je taj odabir dizajna bitan. Lako je pokazati da alat veštačke inteligencije menja ono što kliničar zapisuje. Mnogo je teže — i mnogo važnije — pokazati da menja ono što se događa pacijentu.

Šta su pronašli

Primarni ishod nije se poboljšao. Neuspeh lečenja dogodio se kod 102 od 4,693 pacijenta (2.2%) u AI grupi i 94 od 4,654 pacijenta (2.0%) u kontrolnoj grupi. Neprilagođeni postoci bili su neznatno viši uz AI, ali nakon prilagođavanja procena efekta naginjala je koristi: **prilagođeni odnos izgleda iznosio je 0.77 (95% interval pouzdanosti 0.55 do 1.08, P = 0.13)** — rezultat nije bio statistički značajan. To okretanje smera između sirovih i prilagođenih brojki nije računska greška; prilagođavanje uzima u obzir razlike između klastera kliničara. U oba slučaja interval pouzdanosti bez problema obuhvata „bez efekta", pa se korist ne može tvrditi, a apsolutni je efekat ionako bio vrlo mali.

Za jednostavno objašnjenje kako zajedno čitati odnose šansi, intervale poverenja i P-vrednosti pogledajte vodič za čitanje kliničkih rezultata.

Nije bilo sigurnosnog signala — unutar granica studije. Nijedan ozbiljan štetni događaj nije procenjen kao povezan s alatom, a nezavisna revizija nije našla bezbednosni signal. Autori otvoreno navode granicu tog ohrabrenja: ispitivanje nije imalo dovoljnu statističku snagu za otkrivanje retkih teških šteta i nije imalo unapred određen okvir neinferiornosti ni formalni bezbednosni okvir, pa ne može *dokazati* bezbednost za retke događaje.

Dokumentacija se poboljšala. Među 2,000 susreta koje su pregledali zaslepljeni stručnjaci, kliničari koji su koristili AI Consult imali su bolju kliničku dokumentaciju u svim ocenjivanim područjima — zapisanoj dijagnozi, planu lečenja i ukupnoj potpunosti.

Propisivanje lekova gotovo se nije promenilo. Nije bilo značajne razlike u propisivanju, uključujući pravilnu upotrebu antibiotika (prilagođeni odnos izgleda 0.86, 95% CI 0.48 do 1.55). Alat nije promenio stopu propisivanja antibiotika.

Pacijenti nisu primetili razliku. Među 826 pacijenata koji su ispunili anketu o zadovoljstvu, zadovoljstvo je bilo gotovo jednako u obe grupe, a trajanje konzultacija slično.

Troškovi su blago naginjali naniže. U prilagođenoj analizi troškovi povezani s antibioticima bili su niži u AI grupi — verovatno zbog odabira jeftinijih lekova, a ne manjeg broja recepata — i činilo se da je ušteda po pacijentu na antibioticima veća od troška pokretanja alata po pacijentu. Autori to označavaju kao sugestivan, a ne konačan rezultat: potpuna analiza ukupnog troška vlasništva nije bila deo ispitivanja.

Jednorečenični rezime autora najčišća je verzija: pomoć LLM-a bila je sigurna unutar tih granica, ali nije smanjila neuspeh lečenja unutar 14 dana, a eventualna korist verovatno je skromna.

Zašto „nema značajne razlike" ne znači „ne radi"

Nulti primarni ishod lako je previše protumačiti u bilo kojem smeru. Dve stvari sprečavaju jednostavnu priču.

Prvo, **ispitivanje je bilo projektovano za veći efekat od onoga koji je pronađen.** Ozbiljni loši ishodi

u primarnoj zdravstvenoj zaštiti retki su — ovde oko 2% — pa za razlikovanje malene stvarne koristi od šuma trebaju ogromni uzorci. Naknadni proračuni statističke snage autora sugerišu da bi za otkrivanje efekta veličine kakvu su uočili trebalo **mnogo veće ispitivanje, reda veličine više od 100,000 pacijenata.** Statistički neznačajan rezultat u studiji ove veličine ne isključuje malu stvarnu korist; znači da je ova studija nije mogla razlikovati.

Drugo, **poređenje je delimično bilo zamagljeno.** Greška u konfiguraciji nakratko je nekim kliničarima iz kontrolne grupe omogućila pristup AI Consultu, a kliničari unutar iste mreže razgovaraju i prenose navike preko granice grupa. Oba efekta čine grupe sličnijima i guraju svaku stvarnu razliku prema nuli. Povrh toga, mreža u kojoj je studija sprovedena već je imala relativno visoke standarde, pa je bilo manje prostora da alat pokaže poboljšanje.

Ništa od toga ne spašava naslov o „proboju". Ali znači da ispravno čitanje treba biti odmereno, a ne podrugljivo: na najtežoj i najpoštenijoj krajnjoj tački alat nije dokazivo pomogao pacijentima tokom dve nedelje — dok je merljivo pomogao dokumentaciji i izgledao sigurno.

Šta ovo *ne* dokazuje

- **Ne** pokazuje da je AI poboljšao ishode pacijenata. Na primarnoj 14-dnevnoj krajnjoj tački nije bilo značajne koristi.
- **Ne** pokazuje da je AI beskoristan. Procena efekta išla je u njegovu korist, dokumentacija se poboljšala u svim područjima, a troškovi lekova naginjali su naniže; nulti rezultat kompatibilan je s malom stvarnom koristi koju je ispitivanje bilo premalo da potvrdi.
- **Ne** dokazuje da je alat siguran za retke štete. Nije pronađen bezbednosni signal, ali ispitivanje nije bilo dovoljno veliko niti dizajnirano za potvrđivanje bezbednosti retkih teških događaja.
- **Ne** pokazuje da „AI pobeđuje lečnike" ili ih zamenjuje. Ovo je *podrška* odlučivanju; kliničar je zadržao potpunu ovlast prihvatiti ili odbiti predlog.
- **Ne** generalizuje se automatski. Ispitivanje je sprovedeno u jednoj privatnoj gradskoj mreži u Keniji; ruralna, periurbana i okruženja s višim prihodima mogu se razlikovati u oba smera.
- **Ne** utvrđuje uštede troškova. Signal troškova je sugestivan, a ne završena ekonomska evaluacija.

Koliko su dokazi jaki?

Za centralnu tvrdnju — nema dokazanog smanjenja kratkoročne štete za pacijente — dokaz je snažan *po dizajnu*, a *zaključak* primereno skroman. Prospektivno, unapred registrovano, klasterski randomizovano ispitivanje sa zaslepljenom složenom krajnjom tačkom na nivou pacijenta verovatno je najbolja vrsta dokaza iz stvarnog sveta koju možete prikupiti za ovakav alat. Mnogo je informativnije od rezultata na benčmarku ili studije kliničkih vinjeta.

Za sekundarne nalaze — bolju dokumentaciju, nepromenjeno propisivanje, slično zadovoljstvo i niže troškove antibiotika — dokazi su dobri, ali treba ih čitati kao sekundarne: prateće signale, ne glavnu vest, i podložne istim ograničenjima kontaminacije grupa i jedne zdravstvene mreže.

Za bezbednost su dokazi ohrabrujući, ali ograničeni: signal nije pronađen, ali ovo nije studija osmišljena za nalaženje retkih šteta.

Najkorisniji stav nije ni „radi" ni „nije uspio". On glasi: pažljivo ispitivanje u stvarnom svetu pronašlo je da ovaj AI alat nije pokazao bezbednosni signal i poboljšao je proces nege, ali nije dokazao korist za ishod pacijenta za dve nedelje — a za otkrivanje takve koristi trebalo bi mnogo veće ispitivanje.

Zašto je to važno

Raspravi o medicinskoj veštačkoj inteligenciji upravo ovakvih dokaza hronično nedostaje. Postoje hiljade radova u kojima modeli briljiraju na ispitima i izjednačavaju se s kliničarima na urednim slučajevima. Vrlo je malo velikih, pragmatičnih, randomizovanih ispitivanja koja mere da li su stvarni pacijenti bolje prošli. Ovo je jedno od njih i završava na neglamuroznoj istini: položiti test nije isto što i pomoći pacijentu.

Taj jaz je cela priča. Alat može biti stvarno koristan kliničarima — jasnije beleške, niži računi za lekove, dodatni par očiju — i ipak ne promeniti čvrstu krajnju tačku za pacijente tokom dve nedelje. Obe stvari mogu istovremeno biti tačne, a zreo zdravstveni sistem mora ih držati zajedno umesto da izabere onu koja mu više odgovara.

To takođe korisno pomiče teret dokazivanja. Ako kompanija želi tvrditi da njena klinička veštačka inteligencija poboljšava nega, relevantan dokaz nije skala rezultata. To je ispitivanje poput ovoga, na ishodima koji su važni pacijentima — i, idealno, veće ispitivanje, jer poštena lekcija ovde glasi da su za otkrivanje skromnih koristi potrebne velike studije.

Kratak rezime

Pragmatično, klasterski randomizovano ispitivanje u 16 kenijskih ustanova primarne zdravstvene zaštite testiralo je alat za podršku odlučivanju zasnovan na generativnoj veštačkoj inteligenciji („AI Consult"), dodan elektroničkom kartonu kojim su se koristili klinički službenici. Među 9,691 pacijenata, stručno procenjeni složeni ishod neuspeha lečenja unutar 14 dana iznosio je 2.2% uz alat i 2.0% bez njega (prilagođeni odnos izgleda 0.77, 95% CI 0.55–1.08, P = 0.13) — bez značajne razlike. Alat nije pokazao bezbednosni signal, poboljšao je kliničku dokumentaciju u svim ocenjivanim područjima, nije promenio propisivanje, nije promenio zadovoljstvo pacijenata i bio je povezan s nešto nižim troškovima antibiotika. Autori zaključuju da je bio siguran unutar tih granica, ali nije smanjio neuspeh lečenja, pri čemu je eventualna korist verovatno skromna; za otkrivanje efekta uočene veličine trebalo bi mnogo veće ispitivanje (reda veličine 100,000 pacijenata). Rezultat ne pokazuje da klinički AI poboljšava ishode pacijenata, niti da je beskoristan — pokazuje da alat koji je izgledao sigurno i pomagao procesu nije dokazivo pomogao pacijentima za dve nedelje, u jednoj privatnoj gradskoj mreži.

Provera bez ulepšavanja

Šta rad pokazuje: U randomizovanom ispitivanju iz stvarnog sveta dodavanje alata za podršku odlučivanju zasnovanog na LLM-u u primarnu negu nije pokazalo bezbednosni signal, poboljšalo je kvalitet kliničke dokumentacije, nije promenilo propisivanje i nije statistički značajno smanjilo strogi 14-dnevni složeni ishod neuspeha lečenja.

Šta je verovatno, ali nije dokazano: Da alat donosi malo stvarno smanjenje neuspeha lečenja koje je premalo da bi ga ovo ispitivanje otkrilo; da štedi novac kada se uračunaju svi troškovi; da bolja dokumentacija s vremenom dovodi do bolje nezi.

Šta ne pokazuje: Da klinička veštačka inteligencija poboljšava ishode pacijenata; da je sigurna za retke događaje; da zamenjuje ili nadmašuje kliničare; da se rezultati prenose u ruralna ili bogatija okruženja; da je ušteda troškova utvrđena.

Glavna ograničenja: Studija je imala snagu za veći efekat od uočenog (retki ishodi traže vrlo velike uzorke); greška konfiguracije kontaminirala je kontrolnu grupu, a zajedničke navike unutar mreže zamagljuju uporedbu, pri čemu oba efekta guraju rezultat prema nuli; jedna privatna urbana mreža s već visokim standardima; bez unapred određenog okvira neinferiornosti ili bezbednosti; kratko period od 14 dana.

Koliko poverenja treba imati opšti čitalac? Visoko da alat u ovom ispitivanju nije pokazao bezbednosni signal i da je poboljšao dokumentaciju. Visoko da ovde nije dokazivo poboljšao ishode pacijenata nakon 14 dana. Nisko za bilo kakvu tvrdnju da „radi" ili „ne radi" kao intervencija koja koristi pacijentima — to je pitanje zaista nerešeno i treba mnogo veću studiju. Primeren stav: pažljiv rezultat iz stvarnog sveta o alatu koji nije pokazao bezbednosni signal i poboljšao je proces nege, a ne presuda da AI preobražava — ili uništava — primarnu zdravstvenu zaštitu.

Izvori

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