

The Clean Paper

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

Klinička vještačka inteligencija koja je djelovala sigurno i poboljšala dokumentaciju — ali nije poboljšala ishode pacijenata

Pragmatično, klasterski randomizirano ispitivanje u 16 kenijskih ustanova primarne zdravstvene zaštite testiralo je alat za podršku odlučivanju zasnovan na generativnoj vještačkoj inteligenciji, dodan kartonu koji koriste klinički službenici. Među 9,691 pacijentom, strogi složeni ishod neuspjeha liječenja nakon 14 dana, koji su procjenjivali stručnjaci, iznosio je 2.2% uz alat i 2.0% bez njega (prilagođeni omjer izgleda 0.77, 95% CI 0.55–1.08, P = 0.13) — bez značajne razlike. Alat nije pokazao sigurnosni signal, poboljšao je dokumentaciju u svim ocijenjenim područjima, nije promijenio propisivanje lijekova ni zadovoljstvo i pokazao je trend prema nižim troškovima antibiotika. To nije neuspjela studija; to je rijetka, poštena studija — mjerena na pacijentima, a ne na benchmarkovima — koja završava neglamuroznom istinom da alat koji je djelovao sigurno i pomagao procesu nije dokazivo pomogao pacijentima u roku od dvije sedmice te da bi za otkrivanje eventualne koristi bilo potrebno 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

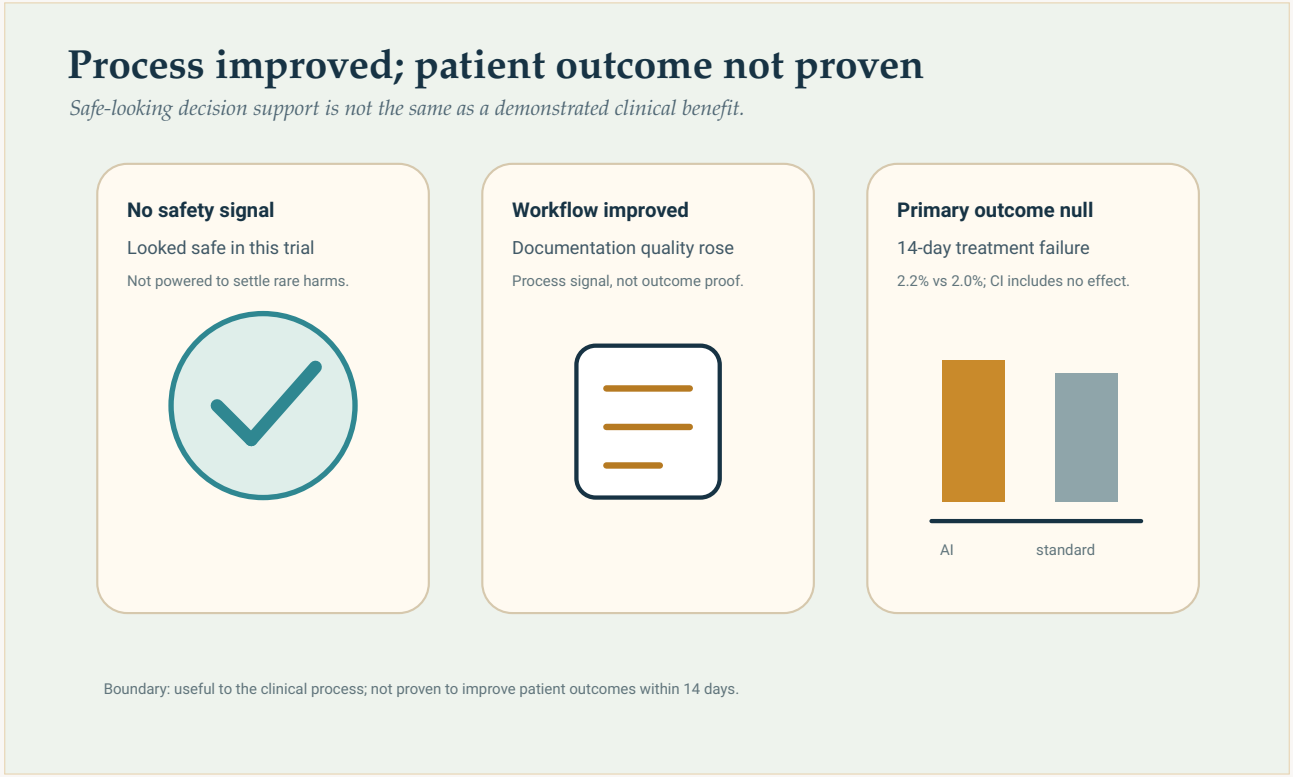
Sigurno i korisno nije isto što i efikasno

Većina naslova o medicinskoj vještačkoj inteligenciji temelji se na pogrešnoj vrsti studije. Model dobro riješi ispitna pitanja ili nadmaši ljekare na pažljivo odabranim kliničkim vinjetama i priča se napiše sama: mašina je spreman za ambulantu.

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

Pošten odgovor je ne — barem ne mjerljivo i ne unutar 14 dana. Alat nije pokazao sigurnosni signal. Poboljšao je kvalitet kliničke dokumentacije. Možda je čak smanjio dio troškova lijekova. Ali nije statistički značajno smanjio neuspjehe liječenja, a autori oprezno navode da je eventualna korist vjerovatno skromna.

To nije neuspjeh studije. To je studija koja radi svoj posao. Ovako izgleda odgovorno prikupljanje dokaza o vještačkoj inteligenciji u medicini kada se mjeri na pacijentima, a ne na benchmarkovima.



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

The Clean Paper CC BY 4.0

Šta su autori uradili

Tim je proveo pragmatično, klasterski randomizirano ispitivanje u **16 ustanova primarne zdravstvene zaštite** privatne zdravstvene mreže Penda Health u okruzima Nairobi i Kiambu u Keniji. Skrb u tim ustanovama uglavnom pružaju **clinical officers** — zdravstveni radnici srednje razine s trogodišnjim stručnim obrazovanjem — često bez lakog pristupa savjetovanju sa starijim kolegama.

Jedinica randomizacije bio je kliničar, a ne pacijent. Randomizirana su **103 clinical officers**: 52 u intervencijsku i 51 u kontrolnu grupu. Obje grupe koristile su isti elektronički 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 liječenja. Kliničari su zadržali potpunu autonomiju: mogli su prihvatiti, izmijeniti ili zanemariti prijedloge.

> **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 elektronički karton (EasyClinicov EMR) i vođen sistemskim promptovima napisanima tako da slijede kenijske nacionalne smjernice liječenja; autori su objavili cijeli instrukcijski 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" općenito. Autori naglašavaju isto: svoj rezultat nazivaju *vremenskim benchmarkom, a ne fiksnom procjenom sposobnosti*. Noviji model, drukčiji prompt ili manje digitalizirana ambulanta mogli bi promijeniti ishod.
> Što se tiče nezavisnosti: OpenAI je naknadno pružio nenovčanu podršku (kredite za računarske resurse u oblaku i tehničke savjete za korištenje API-ja), ali autori navode da je odluka o korištenju OpenAI-ja donesena *prije* 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 namjerno je bio strogo usmjeren na pacijente**: stručno procijenjen *složeni ishod neuspjeha liječenja* unutar 14 dana od posjeta — panel kliničara, zaslijepljen za grupu kojoj je pacijent pripadao, procjenjivao je je li pacijent imao loš ishod poput neriješene ili pogoršane bolesti. Ispitivanje je unaprijed registrirano (Pan-African Clinical Trials Registry 202502499779176).

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

Šta su pronašli

Primarni ishod nije se poboljšao. Neuspjeh liječenja 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 prilagodbe procjena učinka naginjala je koristi: **prilagođeni omjer izgleda iznosio je 0.77 (95% interval pouzdanosti 0.55 do 1.08, P = 0.13)** — rezultat nije bio statistički značajan. To okretanje smjera između sirovih i prilagođenih brojki nije računska greška; prilagodba uzima u obzir razlike između klastera kliničara. U oba slučaja interval pouzdanosti bez problema obuhvata „bez učinka", pa se korist ne može tvrditi, a apsolutni je učinak ionako bio vrlo malen.

Za jednostavno objašnjenje kako zajedno čitati omjere izgleda, intervale pouzdanosti i P-vrijednosti pogledajte vodič za čitanje kliničkih rezultata.

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

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

Propisivanje lijekova gotovo se nije promijenilo. Nije bilo značajne razlike u propisivanju, uključujući pravilnu upotrebu antibiotika (prilagođeni omjer izgleda 0.86, 95% CI 0.48 do 1.55). Alat nije promijenio stopu propisivanja antibiotika.

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

Troškovi su blago naginjali prema dolje. U prilagođenoj analizi troškovi povezani s antibioticima bili su niži u AI grupi — vjerovatno zbog odabira jeftinijih lijekova, 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 dio ispitivanja.

Jednorečenični sažetak autora najčišća je verzija: pomoć LLM-a bila je sigurna unutar tih granica, ali nije smanjila neuspjeh liječenja unutar 14 dana, a eventualna korist vjerovatno je skromna.

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

Nulti primarni ishod lako je previše protumačiti u bilo kojem smjeru. Dvije stvari sprečavaju jednostavnu priču.

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

u primarnoj zdravstvenoj zaštiti rijetki su — ovdje oko 2% — pa za razlikovanje malene stvarne koristi od šuma trebaju ogromni uzorci. Naknadni izračuni statističke snage autora sugeriraju da bi za otkrivanje učinka veličine kakvu su opazili 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 razlučiti.

Drugo, **uporedba je djelomično bila zamagljena.** 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 učinka čine grupe sličnijima i guraju svaku stvarnu razliku prema nuli. Povrh toga, mreža u kojoj je studija provedena 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 odmjereno, a ne podrugljivo: na najtežoj i najpoštenijoj krajnjoj tački alat nije dokazivo pomogao pacijentima tokom dva sedmice — dok je mjerljivo 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. Procjena učinka išla je u njegovu korist, dokumentacija se poboljšala u svim područjima, a troškovi lijekova naginjali su prema dolje; nulti rezultat kompatibilan je s malom stvarnom koristi koju je ispitivanje bilo premalo da potvrdi.
- **Ne** dokazuje da je alat siguran za rijetke štete. Nije pronađen sigurnosni signal, ali ispitivanje nije bilo dovoljno veliko niti dizajnirano za potvrđivanje sigurnosti rijetkih teških događaja.
- **Ne** pokazuje da „AI pobjeđuje ljekare" ili ih zamjenjuje. Ovo je *podrška* odlučivanju; kliničar je zadržao potpunu ovlast prihvatiti ili odbiti prijedlog.
- **Ne** generalizira se automatski. Ispitivanje je provedeno u jednoj privatnoj urbanoj mreži u Keniji; ruralna, periurbana i okruženja s višim prihodima mogu se razlikovati u oba smjera.
- **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* primjereno skroman. Prospektivno, unaprijed registrirano, klasterski randomizirano ispitivanje sa zaslijepljenom složenom krajnjom tačkom na razini pacijenta približno je najbolja vrsta dokaza iz stvarnog svijeta koju možete prikupiti za ovakav alat. Mnogo je informativnije od rezultata na benchmarku ili studije kliničkih vinjeta.

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

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

Najkorisniji stav nije ni „radi" ni „nije uspio". On glasi: pažljivo ispitivanje u stvarnom svijetu pronašlo je da ovaj AI alat nije pokazao sigurnosni signal i poboljšao je proces zdravstvene zaštite, ali nije dokazao korist za ishod pacijenta u dvije sedmice — a za otkrivanje takve koristi trebalo bi mnogo veće ispitivanje.

Zašto je to važno

Raspravi o medicinskoj vještačkoj inteligenciji upravo ovakvih dokaza kronično nedostaje. Postoje hiljade radova u kojima modeli briljiraju na ispitima i izjednačavaju se s kliničarima na urednim slučajevima. Vrlo je malo velikih, pragmatičnih, randomiziranih ispitivanja koja mjere jesu li 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 cijela priča. Alat može biti stvarno koristan kliničarima — jasnije bilješke, niži računi za lijekove, dodatni par očiju — i ipak ne promijeniti čvrstu krajnju tačku za pacijente tokom dva sedmice. Obje stvari mogu istovremeno biti tačne, a zreo zdravstveni sistem mora ih držati zajedno umjesto da izabere onu koja mu više odgovara.

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

Kratak sažetak

Pragmatično, klasterski randomizirano ispitivanje u 16 kenijskih ustanova primarne zdravstvene zaštite testiralo je alat za podršku odlučivanju temeljen na generativnoj vještačkoj inteligenciji („AI Consult"), dodan elektroničkom kartonu kojim su se koristili clinical officers. Među 9,691 pacijenata, stručno procijenjeni složeni ishod neuspjeha liječenja unutar 14 dana iznosio je 2.2% uz alat i 2.0% bez njega (prilagođeni omjer izgleda 0.77, 95% CI 0.55–1.08, P = 0.13) — bez značajne razlike. Alat nije pokazao sigurnosni signal, poboljšao je kliničku dokumentaciju u svim ocjenjivanim područjima, nije promijenio propisivanje, nije promijenio zadovoljstvo pacijenata i bio je povezan s nešto nižim troškovima antibiotika. Autori zaključuju da je bio siguran unutar tih granica, ali nije smanjio neuspjeh liječenja, pri čemu je eventualna korist vjerovatno skromna; za otkrivanje učinka opažene 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 u dvije sedmice, u jednoj privatnoj urbanoj mreži.

Provjera bez uljepšavanja

Šta rad pokazuje: U randomiziranom ispitivanju iz stvarnog svijeta dodavanje alata za podršku odlučivanju temeljenog na LLM-u u primarnu zdravstvenu zaštitu nije pokazalo sigurnosni signal, poboljšalo je kvalitet kliničke dokumentacije, nije promijenilo propisivanje i nije statistički značajno smanjilo strogi 14-dnevni složeni ishod neuspjeha liječenja.

Šta je vjerovatno, ali nije dokazano: Da alat donosi malo stvarno smanjenje neuspjeha liječenja 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 zdravstvene zaštite.

Šta ne pokazuje: Da klinička vještačka inteligencija poboljšava ishode pacijenata; da je sigurna za rijetke događaje; da zamjenjuje 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 učinak od opaženog (rijetki ishodi traže vrlo velike uzorke); greška konfiguracije kontaminirala je kontrolnu grupu, a zajedničke navike unutar mreže zamagljuju uporedbu, pri čemu oba učinka guraju rezultat prema nuli; jedna privatna urbana mreža s već visokim standardima; bez unaprijed određenog okvira neinferiornosti ili sigurnosti; kratko period od 14 dana.

Koliko povjerenja treba imati opći čitalac? Visoko da alat u ovom ispitivanju nije pokazao sigurnosni signal i da je poboljšao dokumentaciju. Visoko da ovdje 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 neriješeno i treba mnogo veću studiju. Primjeren stav: pažljiv rezultat iz stvarnog svijeta o alatu koji nije pokazao sigurnosni signal i poboljšao je proces zdravstvene zaštite, 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