The Clean Paper

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

Klinična umetna inteligenca, ki je delovala varno in izboljšala dokumentacijo — vendar ni izboljšala izidov bolnikov

Pragmatično, grozdno randomizirano preskušanje v 16 kenijskih ambulantah primarnega zdravstva je preizkusilo generativno orodje za podporo odločanju, dodano elektronskemu zapisu, ki ga uporabljajo klinični uradniki. Med 9.691 bolniki je strogi, strokovno presojeni sestavljeni 14-dnevni izid neuspeha zdravljenja znašal 2,2 % z orodjem proti 2,0 % brez njega (prilagojeno razmerje obetov 0,77; 95-% IZ 0,55–1,08; P = 0,13) — brez statistično značilne razlike. Orodje ni pokazalo varnostnega signala, izboljšalo je dokumentacijo v vseh ocenjenih domenah, ni spremenilo predpisovanja ali zadovoljstva in je nakazovalo nižje stroške antibiotikov. To ni neuspešna študija; je redka poštena študija, merjena na bolnikih in ne na benchmarkih — ter pristane pri nebleščeči resnici, da orodje, ki je delovalo varno in pomagalo procesu, v dveh tednih ni dokazljivo pomagalo bolnikom, za zaznavo morebitne majhne koristi pa bi bilo potrebno veliko večje preskušanje.

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

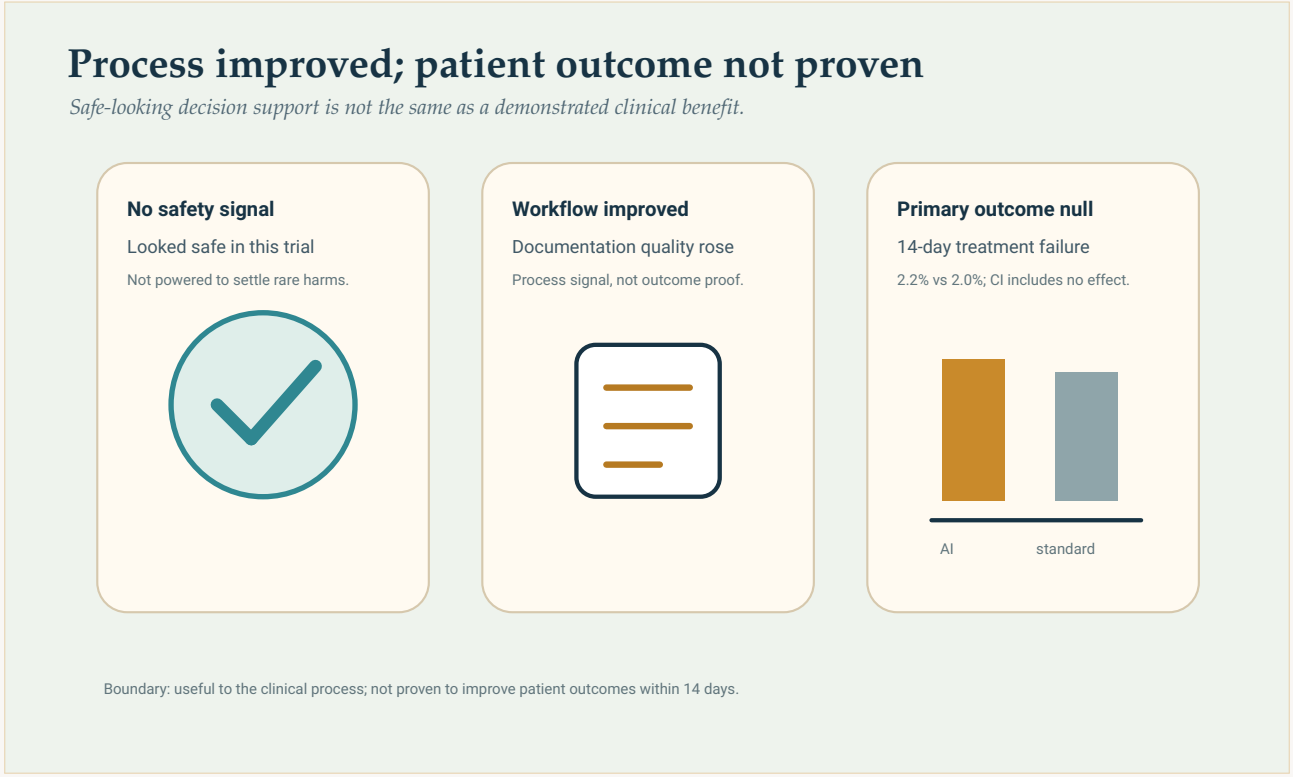
Varno in koristno ni isto kot učinkovito

Večina naslovov o medicinski umetni inteligenci temelji na napačni vrsti raziskave. Model dobro rešuje izpitna vprašanja ali premaga zdravnike na skrbno izbranih kliničnih vinjetah in zgodba se napiše sama: stroj je pripravljen za ambulanto.

To preskušanje je naredilo nekaj težjega in redkejšega. Generativno orodje za podporo odločanju je postavilo v pravo primarno zdravstvo, v prave ambulante, k pravim bolnikom in zastavilo vprašanje, ki dejansko šteje: so bili bolniki na koncu na boljšem?

Pošten odgovor je ne — vsaj ne merljivo in ne v 14 dneh. Orodje ni pokazalo varnostnega signala. Izboljšalo je kakovost klinične dokumentacije. Morda je celo nekoliko znižalo stroške nekaterih zdravil. Ni pa statistično značilno zmanjšalo neuspehov zdravljenja, avtorji pa previdno poudarjajo, da je morebitna korist verjetno majhna.

To ni neuspeh študije. To je študija, ki deluje, kot mora. Tako je videti odgovoren dokaz o umetni inteligenci v medicini, ko ga merimo pri bolnikih namesto na benchmarkih.



Orodje ni pokazalo varnostnega signala in je izboljšalo klinično dokumentacijo, vendar se vnaprej določeni 14-dnevni izid bolnikov ni statistično značilno izboljšal. Pomoč pri procesu ni isto kot dokazana korist za bolnika.

Kaj so naredili avtorji

Ekipa je izvedla pragmatično, grozdno randomizirano preskušanje v **16 ambulantah primarnega zdravstva** zasebne zdravstvene mreže Penda Health v okrožjih Nairobi in Kiambu v Keniji. Večino oskrbe tam zagotavljajo **clinical officers** — zdravstveni delavci srednje ravni s triletnim strokovnim izobraževanjem — pogosto brez enostavnega dostopa do starejšega specialista za posvet.

Randomiziran je bil klinik, ne bolnik. Naključno so razporedili **103 klinične uradnike**: 52 v intervencijsko in 51 v kontrolno skupino. Obe skupini sta uporabljali isti elektronski zdravstveni zapis v oblaku. Intervencijska skupina je imela dodatno **AI Consult** (različica 2.0), orodje za podporo odločanju, zgrajeno na velikem jezikovnem modelu **OpenAI GPT-4o** in vgrajeno v zapis. Prebralo je podatke, ki jih je klinik vnesel, ter lahko opozorilo na možne težave z diagnozo ali načrtom zdravljenja. Kliniki so ohranili popolno avtonomijo: predlog so lahko sprejeli, spremenili ali prezrli.

> **Kateri model je bil uporabljen in zakaj podrobnosti štejejo**
>
> Orodje je bilo **AI Consult 2.0** z modelom **OpenAI GPT-4o** (izdaja maja 2025), dostopnim prek komercialnega API-ja OpenAI z enterprise licenco in nastavljenim na nizko naključnost (temperature 0,1). Delovalo je znotraj prilagojenega elektronskega zapisa EasyClinic ter ga usmerjalo sistemsko navodilo, napisano skladno s kenijskimi nacionalnimi smernicami zdravljenja; avtorji so objavili celoten prompt.
> Zakaj toliko podrobnosti? Ker rezultat velja za *en določen sistem* — eno različico modela, en prompt, en zapis in eno nastavitev — ne za »LLM v medicini« na splošno. Avtorji poudarjajo isto: rezultat imenujejo *časovni benchmark, ne fiksna ocena zmogljivosti*. Novejši model, drugačen prompt ali manj digitalizirana ambulanta bi lahko izid spremenili.
> Kar zadeva neodvisnost: OpenAI je pozneje prispeval podporo v naravi (dobropise za računske vire v oblaku in tehnične napotke za uporabo API-ja), vendar avtorji navajajo, da je bila odločitev za OpenAI sprejeta *pred* to ponudbo in da OpenAI ni sodeloval pri zasnovi preskušanja, zbiranju ali analizi podatkov niti pri odločitvi za objavo.

Med 22. aprilom in 16. julijem 2025 je bilo vključenih **9.691 bolnikov**. **Primarni izid je bil namenoma strog in usmerjen v bolnika**: strokovno presojeni *sestavljeni izid neuspeha zdravljenja* v 14 dneh po obisku. Komisija klinikov, slepa za skupino, je ocenjevala, ali je bolnik imel slab izid, na primer nerešeno ali poslabšano bolezen. Preskušanje je bilo vnaprej registrirano (Pan-African Clinical Trials Registry 202502499779176).

Prav ta izbira zasnove je bistvena. Preprosto je pokazati, da AI spremeni, kaj klinik zapiše. Veliko težje — in veliko pomembneje — je pokazati, da spremeni, kaj se zgodi z bolnikom.

Kaj so ugotovili

Primarni izid se ni izboljšal. Neuspeh zdravljenja se je pojavil pri 102 od 4.693 bolnikov (2,2 %) v skupini z AI in pri 94 od 4.654 (2,0 %) v kontrolni skupini. Surova odstotka sta bila malenkost

višja z AI, po prilagoditvi pa je točkovna ocena nakazovala korist: **prilagojeno razmerje obetov 0,77 (95-% interval zaupanja 0,55–1,08; P = 0,13)** — brez statistične značilnosti. Obrat med surovimi in prilagojenimi številkami ni računska napaka; prilagoditev upošteva razlike med grozdi klinikov. V vsakem primeru interval zaupanja udobno vključuje »brez učinka«, zato koristi ni mogoče trditi, absolutni učinek pa je bil zelo majhen.

Za preprost način skupnega branja razmerja obetov, intervala zaupanja in vrednosti P glejte vodnik za branje kliničnega rezultata.

Brez varnostnega signala — z omejitvami. Noben resen neželeni dogodek ni bil ocenjen kot povezan z orodjem, neodvisni pregled pa ni našel varnostnega signala. Avtorji jasno omejujejo, koliko pomiritve to prinaša: preskušanje ni imelo dovolj moči za zaznavanje redkih hudih škod in ni imelo vnaprej določenega okvira neinferiornosti ali formalnega varnostnega načrta, zato varnosti za redke dogodke ne more *dokazati*.

Dokumentacija se je izboljšala. Med 2.000 obiski, ki so jih pregledali slepi strokovnjaki, so kliniki z AI Consult ustvarili boljšo klinično dokumentacijo v vseh ocenjenih domenah — zapisu diagnoze, načrtu zdravljenja in splošni popolnosti.

Predpisovanje se skoraj ni spremenilo. Statistično značilne razlike pri predpisovanju ni bilo, tudi pri pravilni uporabi antibiotikov ne (prilagojeno razmerje obetov 0,86; 95-% IZ 0,48–1,55). Orodje ni spremenilo stopnje predpisovanja antibiotikov.

Bolniki razlike niso opazili. Med 826 bolniki, ki so izpolnili vprašalnik zadovoljstva, je bilo zadovoljstvo v obeh skupinah praktično enako, podobno dolgi so bili tudi posveti.

Stroški so nakazovali rahlo zmanjšanje. V prilagojeni analizi so bili stroški, povezani z antibiotiki, v skupini z AI nižji — verjetno zaradi cenejših izbir, ne zaradi manj receptov — prihranek antibiotikov na bolnika pa je bil videti večji od stroška uporabe orodja na bolnika. Avtorji to označujejo kot namig, ne kot dokončen rezultat: celovita analiza vseh stroškov lastništva ni bila del preskušanja.

Enovrstični povzetek avtorjev je najčistejši: pomoč LLM je bila v teh mejah videti varna, vendar v 14 dneh ni zmanjšala neuspeha zdravljenja, morebitna korist pa je verjetno majhna.

Zakaj »brez značilne razlike« ne pomeni »ne deluje«

Ničelni primarni izid je lahko prehitro interpretirati v obe smeri. Preprosto zgodbo preprečujeta dve stvari.

Prvič, **preskušanje je bilo zasnovano za večji učinek od opaženega.** Resni slabi izidi v primarnem zdravstvu so redki — tukaj okoli 2 % — zato so za ločevanje majhne resnične koristi od šuma potrebne ogromne številke. Post-hoc izračuni moči avtorjev kažejo, da bi za zaznavo učinka velikosti, ki so ga opazili, potrebovali **bistveno večje preskušanje, reda več kot 100.000 bolnikov**. Neznačilen rezultat v tej velikosti študije ne izključuje majhne resnične koristi; pomeni, da je ta študija ni mogla razločiti.

Drugič, **primerjava se je deloma zabrisala.** Zaradi konfiguracijske napake so nekateri kliniki v kon-

trolni skupini kratek čas dobili dostop do AI Consult, kliniki v isti mreži pa se pogovarjajo in prenašajo navade čez mejo med skupinama. Oboje skupini naredi bolj podobni in morebitno resnično razliko potisne proti ničli. Poleg tega je gostiteljska mreža že delovala po razmeroma visokih standardih, zato je bilo manj prostora za izboljšanje.

Nič od tega ne rešuje naslova o »preboju«. Pomeni pa, da mora biti interpretacija umerjena, ne preprosto negativna: na najtežjem in najbolj poštenem izidu orodje v dveh tednih ni dokazljivo pomagalo bolnikom — hkrati pa je merljivo pomagalo pri dokumentaciji in ni pokazalo varnostnega signala.

Česa to ne dokazuje

- **Ne** kaže, da je AI izboljšala izide bolnikov. Pri primarnem 14-dnevnem izidu ni bilo značilne koristi.
- **Ne** kaže, da je AI neuporabna. Točkovna ocena ji je bila naklonjena, dokumentacija se je izboljšala, stroški zdravil so se nagibali navzdol; ničelni rezultat je združljiv z majhno resnično koristjo, ki je bilo preskušanje premajhno, da bi jo potrdilo.
- **Ne** dokazuje varnosti pri redkih škodah. Varnostnega signala ni bilo, vendar študija ni imela moči ali zasnove za potrjevanje redkih hudih dogodkov.
- **Ne** kaže, da »AI premaga zdravnike« ali jih nadomešča. To je *podpora* odločanju; klinik je obdržal polno pristojnost za sprejem ali zavrnitev predloga.
- **Ne** posplošuje samodejno. Preskušanje je potekalo v eni zasebni urbani mreži v Keniji; podeželska, periurbana ali bogatejša okolja se lahko razlikujejo v katerokoli smer.
- **Ne** dokazuje prihranka stroškov. Signal stroškov je namig, ne končana ekonomska analiza.

Kako močni so dokazi?

Za osrednjo trditev — da ni dokazanega zmanjšanja kratkoročne škode bolnikov — je dokazna zasnova močna, zaključek pa ustrezno skromen. Prospektivno, vnaprej registrirano, grozdno randomizirano preskušanje s slepo presojanim sestavljenim izidom na ravni bolnika je skoraj najboljši realni dokaz, ki ga lahko zberemo za takšno orodje. Je bistveno bolj informativno od benchmarka ali študije kliničnih vinjet.

Za sekundarne rezultate — boljšo dokumentacijo, nespremenjeno predpisovanje, podobno zadovoljstvo, nižje stroške antibiotikov — so dokazi dobri, vendar jih je treba brati kot sekundarne: podporni signali, ne naslov, in podvrženi istim omejitvam kontaminacije in ene same mreže.

Za varnost so dokazi pomirjujoči, vendar omejeni: signala niso našli, vendar študija ni bila zasnovana za redke škode.

Najuporabnejše stališče ni niti »deluje« niti »ni uspelo«. Je: skrbno realno preskušanje je ugotovilo, da ta AI ni pokazala varnostnega signala in je izboljšala proces oskrbe, ni pa v dveh tednih dokazala koristi pri izidu bolnikov — za zaznavo morebitne majhne koristi pa bi bila potrebna precej večja

študija.

Zakaj je pomembno

Razpravi o medicinski AI primanjkuje prav takšnih dokazov. Na tisoče člankov kaže modele, ki blestijo na izpitih in se na urejenih primerih merijo z zdravniki. Velikih pragmatičnih randomiziranih preskušanj, ki merijo, ali je resničnim bolnikom bolje, je zelo malo. To je eno izmed njih in pristane pri nebleščeči resnici: opraviti test ni isto kot pomagati bolniku.

Ta vrzel je celotna zgodba. Orodje je lahko klinikom resnično koristno — jasnejši zapisi, nižji račun za zdravila, dodatni par oči — in kljub temu v štirinajstih dneh ne premakne trdega izida bolnikov. Oboje je lahko res hkrati in zrel zdravstveni sistem mora držati obe dejstvi skupaj, namesto da izbere priročnejšega.

Študija tudi koristno prestavi dokazno breme. Če podjetje želi trditi, da njegova klinična AI izboljšuje oskrbo, relevanten dokaz ni leaderboard. Je preskušanje, kakršno je to, z izidi, ki so pomembni bolnikom — in po možnosti še večje, saj je poštena lekcija tukaj, da so za majhne koristi potrebne velike študije.

Kratek povzetek

Pragmatično, grozdno randomizirano preskušanje v 16 kenijskih ambulantah primarnega zdravstva je preizkusilo generativno orodje za podporo odločanju AI Consult, dodano elektronskemu zapisu kliničnih uradnikov. Med 9.691 bolniki je strokovno presojeni sestavljeni neuspeh zdravljenja v 14 dneh znašal 2,2 % z orodjem in 2,0 % brez njega (prilagojeno razmerje obetov 0,77; 95-% IZ 0,55–1,08; P = 0,13) — brez statistično značilne razlike. Orodje ni pokazalo varnostnega signala, izboljšalo je klinično dokumentacijo v vseh ocenjenih domenah, ni spremenilo predpisovanja, zadovoljstvo bolnikov je ostalo enako, povezano pa je bilo z nekoliko nižjimi stroški antibiotikov. Avtorji sklepajo, da je bilo v teh mejah videti varno, vendar ni zmanjšalo neuspeha zdravljenja; morebitna korist je verjetno majhna in bi za zaznavo učinka opažene velikosti potrebovali precej večje preskušanje, reda 100.000 bolnikov. Rezultat ne kaže, da klinična AI izboljšuje izide bolnikov, niti da je neuporabna — kaže, da orodje, ki je delovalo varno in pomagalo procesu, v dveh tednih ni dokazljivo pomagalo bolnikom v eni zasebni urbani mreži.

Preverjanje brez olepševanja

Kaj članek pokaže: V realnem randomiziranem preskušanju dodatek LLM-orodja za podporo odločanju v primarni zdravstveni zapis ni pokazal varnostnega signala, izboljšal je kakovost klinične dokumentacije, ni spremenil predpisovanja in ni statistično značilno zmanjšal strogega 14-dnevnega

sestavljenega neuspeha zdravljenja.

Kaj je verjetno, vendar ne dokazano: Da orodje povzroči majhno resnično zmanjšanje neuspeha zdravljenja, premajhno za zaznavo v tej študiji; da prihrani denar, ko se vključijo vsi stroški; da boljša dokumentacija sčasoma vodi v boljšo oskrbo.

Česa ne pokaže: Da klinična AI izboljša izide bolnikov; da je varna pri redkih dogodkih; da nadomešča ali presega klinike; da rezultati veljajo v podeželskih ali bogatejših okoljih; da so prihranki dokazani.

Glavne omejitve: Študija je imela moč za večji učinek od opaženega; redki izidi zahtevajo zelo velike vzorce. Konfiguracijska napaka je deloma kontaminirala kontrolno skupino, skupna mreža pa je zabrisala primerjavo — oboje potiska razliko proti ničli. Ena zasebna urbana mreža z že visokimi standardi; brez vnaprej določenega okvira neinferiornosti ali varnosti; kratko 14-dnevno obdobje.

Koliko zaupanja naj ima splošni bralec? Visoko, da orodje v tem preskušanju ni pokazalo varnostnega signala in je izboljšalo dokumentacijo. Visoko tudi, da tukaj ni dokazljivo izboljšalo 14-dnevnih izidov bolnikov. Nizko za trditev, da kot intervencija za korist bolnika »deluje« ali »ne deluje« — to vprašanje res ostaja odprto in zahteva bistveno večjo študijo. Primeren sklep: skrbno realno preskušanje orodja, ki ni pokazalo varnostnega signala in je izboljšalo proces oskrbe, ne razsodba, da AI preobraža — ali uničuje — primarno zdravstvo.

Viri

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