The Clean Paper

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

Un AI clinic care a părut sigur și a îmbunătățit documentația — dar nu a îmbunătățit rezultatele pacienților

Un trial pragmatic, cluster-randomizat, în 16 unități de îngrijire primară din Kenya, a testat un instrument de suport decizional cu AI generativă adăugat fișei folosite de clinical officers. Dintre 9.691 de pacienți, compozitul strict de eșec terapeutic la 14 zile, judecat de experți, a fost 2,2% cu instrumentul versus 2,0% fără (odds ratio ajustat 0,77, IC 95% 0,55–1,08, P = 0,13) — nicio diferență semnificativă. Instrumentul nu a arătat semnal de siguranță, a îmbunătățit documentația în toate domeniile evaluate, a lăsat prescrierea și satisfacția neschimbate și a tins spre costuri mai mici cu antibioticele. Acesta nu este un studiu eșuat; este unul rar și onest — măsurat pe pacienți, nu pe benchmarkuri — și aterizează pe adevărul neglamouros că un instrument care a părut sigur și a ajutat procesul nu a ajutat demonstrabil pacienții în două săptămâni, iar detectarea oricărui asemenea beneficiu ar cere un trial mult mai mare.

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

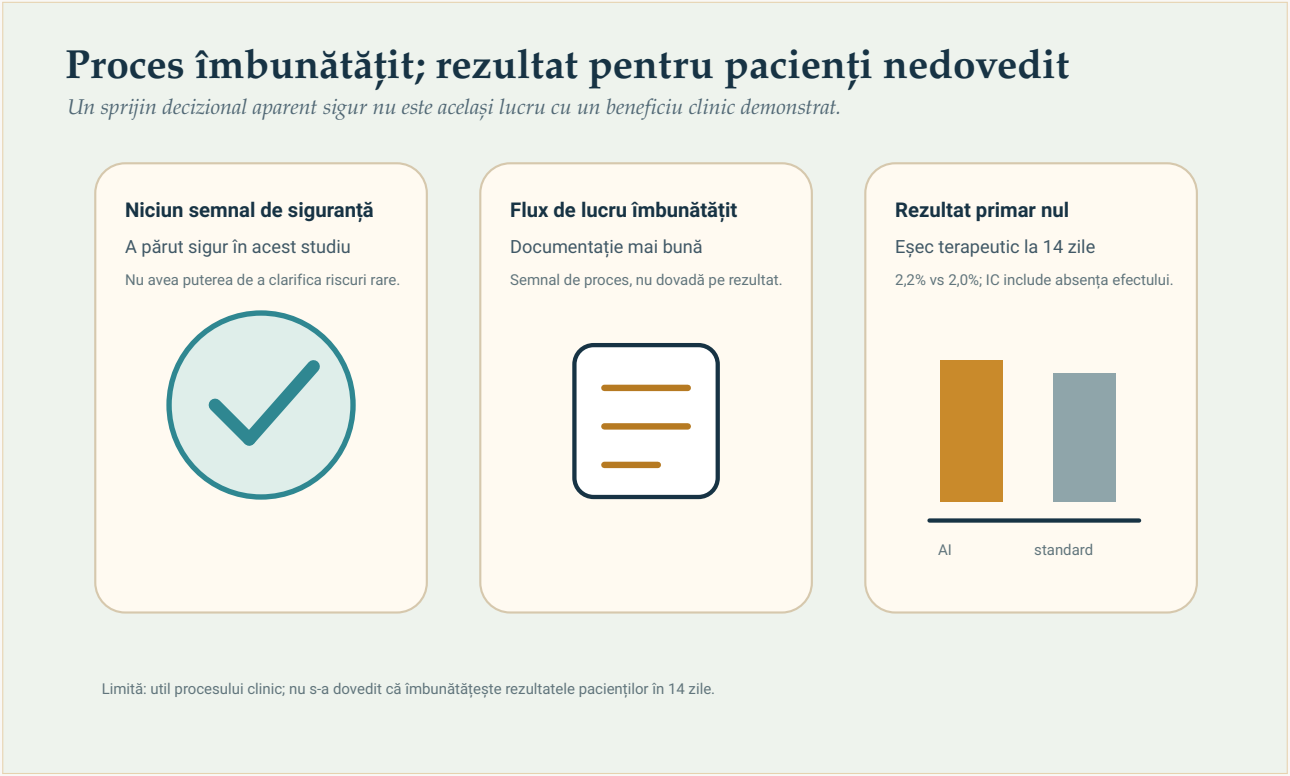
Sigur și util nu înseamnă același lucru cu eficient

Multe titluri despre AI medicală pornesc de la tipul greșit de studiu. Un model obține scoruri bune la întrebări de examen sau îi depășește pe medici în cazuri clinice atent selectate, iar concluzia pare să vină de la sine: sistemul este gata pentru clinică.

Acest studiu a făcut ceva mai dificil și mai rar. A introdus un instrument de suport decizional cu AI generativă în medicina primară reală, în unități reale și cu pacienți reali, apoi a pus întrebarea care contează cu adevărat: au avut pacienții rezultate mai bune?

Răspunsul prudent este nu — cel puțin nu într-un mod măsurabil în 14 zile. Instrumentul nu a evidențiat probleme de siguranță, a îmbunătățit calitatea documentației clinice și este posibil să fi redus unele costuri cu medicamentele. Dar nu a redus semnificativ eșecurile terapeutice, iar autorii subliniază că orice beneficiu, dacă există, este probabil modest.

Acesta nu este un eșec al studiului, ci exact rolul unui astfel de studiu. Așa arată o evaluare responsabilă a AI în medicină atunci când măsoară ce se întâmplă cu pacienții, nu doar performanța pe teste de referință.



Instrumentul nu a evidențiat probleme de siguranță și a îmbunătățit documentația clinică, dar rezultatul prespecificat pentru pacienți la 14 zile nu s-a îmbunătățit semnificativ. Îmbunătățirea procesului nu este același lucru cu un beneficiu demonstrat pentru pacient.

Ce au făcut autorii

Echipa a desfășurat un studiu pragmatic, randomizat pe clustere, în **16 unități de îngrijire primară** operate de o rețea privată de sănătate (Penda Health) în comitatele Nairobi și Kiambu, Kenya. Îngrijirea este oferită în mare parte de **clinical officers** — practicieni de nivel intermediar cu o formare de trei ani — care adesea nu au acces facil la consultarea unui clinician cu mai multă experiență.

Unitatea de randomizare a fost clinicianul, nu pacientul. **103 clinical officers** au fost repartizați aleatoriu: 52 în brațul de intervenție și 51 în cel de control. Ambele grupuri au folosit aceeași fișă medicală electronică în cloud. Grupul de intervenție a avut în plus **„AI Consult”** (versiunea 2.0), un instrument de suport decizional bazat pe modelul lingvistic de mari dimensiuni **GPT-4o de la OpenAI** și integrat în fișa electronică. Instrumentul citea informațiile consemnate de clinician și putea semnala posibile probleme în diagnostic sau în planul de tratament. Clinicienii și-au păstrat autonomia deplină: puteau accepta, modifica sau ignora sugestiile.

> **Ce model era și de ce contează specificul**
>
> Instrumentul era **AI Consult 2.0**, care utiliza **GPT-4o de la OpenAI** (versiunea din mai 2025), accesat prin API-ul comercial OpenAI, cu licență pentru companii și cu un nivel redus de aleatoriu (temperatură 0,1). Era integrat într-o fișă electronică personalizată (EMR-ul EasyClinic) și ghidat de prompturi de sistem concepute pentru a respecta ghidurile naționale de tratament din Kenya; autorii au publicat integral promptul de instrucțiuni.
> De ce contează aceste detalii? Pentru că rezultatul se referă la *un sistem specific* — o anumită versiune de model, un anumit prompt, o anumită fișă electronică și un anumit context — nu la „LLM-uri în medicină” în general. Autorii subliniază același lucru: descriu rezultatul drept un *reper valabil la un anumit moment, nu o estimare fixă a capacității*. Un model mai nou, un prompt diferit sau o clinică mai puțin digitalizată ar putea produce un rezultat diferit.
> În privința independenței: OpenAI a oferit ulterior sprijin în natură (credite pentru calcul în cloud și îndrumare tehnică privind utilizarea API-ului), dar autorii precizează că decizia de a folosi OpenAI fusese luată *înainte* de acea ofertă și că OpenAI nu a avut niciun rol în proiectarea studiului, colectarea sau analiza datelor ori decizia de publicare.

Între 22 aprilie și 16 iulie 2025 au fost înscriși **9.691 de pacienți**. **Rezultatul primar a fost în mod deliberat strict și centrat pe pacient**: un *indicator compozit de eșec terapeutic* în primele 14 zile după consultație. Un grup de clinicieni, care nu știa în ce braț al studiului se afla pacientul, a evaluat dacă apăruse un rezultat negativ, precum persistența sau agravarea bolii. Studiul a fost înregistrat în avans (Pan-African Clinical Trials Registry 202502499779176).

Această alegere metodologică este esențială. Este ușor să arăți că un instrument AI schimbă ceea ce consemnează un clinician. Este mult mai greu — și mult mai relevant — să arăți că schimbă ceea ce i se întâmplă pacientului.

Ce au găsit

Rezultatul primar nu s-a îmbunătățit. Eșecul terapeutic a apărut la 102 din 4.693 de pacienți (2,2%) în brațul AI și la 94 din 4.654 (2,0%) în brațul control. Procentele brute au fost fracțional mai mari cu AI, dar după ajustare estimarea punctuală înclina spre beneficiu: **odds ratio-ul ajustat a fost 0,77 (interval de încredere 95% 0,55–1,08, P = 0,13)** — nu statistic semnificativ. Această inversare între numerele brute și cele ajustate nu este o eroare aritmetică; ajustarea ține cont de diferențele dintre clusterele de clinicieni. Oricum, intervalul de încredere include confortabil „niciun efect", deci nu se poate revendica un beneficiu, iar în termeni absoluți efectul a fost minuscul.

Pentru o explicație în limbaj simplu despre cum se citesc împreună odds ratio, intervalele de încredere și p-value-urile, vezi ghidul pentru rezultate clinice.

Nu au apărut probleme de siguranță — în limitele studiului. Niciun eveniment advers grav nu a fost considerat legat de instrument, iar o evaluare independentă nu a identificat un semnal de siguranță. Autorii precizează însă limita acestei concluzii liniștitoare: studiul nu avea puterea statistică necesară pentru a detecta efecte adverse severe, dar rare, și nu avea un cadru prespecificat de noninferioritate sau de evaluare formală a siguranței. Prin urmare, nu poate *dovedi* siguranța pentru evenimente neobișnuite.

Documentația s-a îmbunătățit. Dintre 2.000 de întâlniri revizuite de experți orbi, clinicienii care foloseau AI Consult au produs documentație clinică mai bună în toate domeniile evaluate — diagnosticul înregistrat, planul de tratament și completitudinea generală.

Prescrierea abia s-a mișcat. Nu a existat diferență semnificativă în prescriere, inclusiv în folosirea corectă a antibioticelor (odds ratio ajustat 0,86, IC 95% 0,48–1,55). Instrumentul nu a schimbat ratele de prescriere a antibioticelor.

Pacienții nu au observat diferență. Dintre 826 de pacienți care au completat un sondaj de satisfacție, satisfacția a fost practic identică între brațe, iar timpii consultațiilor au fost similari.

Costurile au avut o ușoară tendință descendentă. Într-o analiză ajustată, costurile legate de antibiotice au fost mai mici în brațul AI — probabil prin alegerea unor variante mai ieftine, nu prin reducerea numărului de prescripții — iar economia per pacient pentru antibiotice părea să depășească costul utilizării instrumentului. Autorii prezintă însă rezultatul ca sugestiv, nu ca demonstrat: o evaluare completă a costului total de utilizare nu a făcut parte din studiu.

Rezumatul autorilor într-o singură propoziție este și formularea cea mai precisă: asistența bazată pe LLM nu a ridicat probleme de siguranță în limitele studiului, dar nu a redus eșecul terapeutic în 14 zile, iar orice beneficiu este probabil modest.

De ce „nicio diferență semnificativă” nu înseamnă „nu funcționează”

Un rezultat primar nul poate fi ușor suprainterpretat în ambele direcții. Două aspecte împiedică însă o concluzie simplistă.

Mai întâi, **studiul a fost dimensionat pentru a detecta un efect mai mare decât cel observat**. Rezultatele clinice negative grave sunt rare în îngrijirea primară — aproximativ 2% aici — astfel încât distingerea unui beneficiu real, dar mic, de zgomotul statistic necesită eșantioane enorme. Calculele post-hoc ale autorilor sugerează că detectarea unui efect de mărimea celui observat ar necesita un **studiu mult mai mare, cu peste 100.000 de pacienți.** Un rezultat nesemnificativ într-un studiu de această dimensiune nu exclude un beneficiu real mic; înseamnă doar că studiul nu l-a putut detecta.

În al doilea rând, **comparația dintre grupuri a fost parțial estompată**. O eroare de configurare le-a oferit pentru scurt timp unor clinicieni din brațul de control acces la AI Consult, iar clinicienii din aceeași rețea discută între ei și pot adopta practici unii de la alții. Ambele efecte tind să apropie grupurile și să împingă orice diferență reală spre zero. În plus, rețeaua funcționa deja la standarde relativ ridicate, ceea ce lasă mai puțin loc pentru o îmbunătățire vizibilă.

Nimic din toate acestea nu justifică un titlu despre o „descoperire revoluționară”. Dar concluzia corectă trebuie să rămână calibrată, nu categorică: pentru criteriul cel mai relevant pentru pacient, instrumentul nu a demonstrat un beneficiu în primele două săptămâni, deși a îmbunătățit măsurabil documentația și nu a evidențiat probleme de siguranță.

Ce *nu* demonstrează acest studiu

- **Nu** arată că AI a îmbunătățit rezultatele pacienților. Pentru rezultatul primar la 14 zile nu a existat un beneficiu semnificativ.
- **Nu** arată că AI este inutilă. Estimarea punctuală o favoriza, documentația s-a îmbunătățit în toate domeniile evaluate, iar costurile medicamentelor au avut o tendință descendentă; rezultatul nul este compatibil cu un beneficiu real mic pe care studiul nu a avut suficientă putere statistică să-l confirme.
- **Nu** dovedește că instrumentul este sigur pentru daune rare. Nu a arătat semnal de siguranță, dar nu a fost dimensionat sau proiectat să certifice siguranța pentru evenimente severe neobișnuite.
- **Nu** arată că „AI bate doctorii” sau înlocuiește clinicienii. Este suport decizional; clinicianul a păstrat autoritatea completă de a-l accepta sau respinge.
- **Nu** se generalizează automat. Studiul s-a desfășurat într-o singură rețea urbană privată din Kenya; mediile rurale, periurbane și cele cu venituri mai mari ar putea produce rezultate diferite.
- **Nu** stabilește economii de cost. Semnalul de cost este sugestiv, nu o evaluare economică completă.

Cât de solide sunt dovezile?

Pentru afirmația centrală — nicio reducere demonstrată a rezultatelor negative pe termen scurt pentru pacienți — dovezile sunt puternice *prin metodologia studiului,* iar concluzia rămâne pe bună dreptate modestă. Un studiu prospectiv, preînregistrat și randomizat pe clustere, cu un rezultat compozit la nivelul pacientului evaluat în orb, se apropie de cea mai bună dovadă din lumea reală care poate fi obținută pentru un astfel de instrument. Este mult mai informativ decât un scor pe un test de referință sau un studiu pe vignete clinice.

Pentru rezultatele secundare — documentație mai bună, prescriere neschimbată, satisfacție similară și costuri mai mici pentru antibiotice — dovezile sunt bune, dar trebuie interpretate ca atare: rezultate secundare, nu concluzia principală, și supuse acelorași limite legate de contaminarea dintre grupuri și de folosirea unei singure rețele.

În privința siguranței, rezultatele sunt liniștitoare, dar limitate: nu a fost identificat niciun semnal, însă studiul nu a fost dimensionat pentru a detecta efecte adverse rare.

Concluzia cea mai utilă nu este nici „funcționează", nici „a eșuat". Un studiu atent, desfășurat în lumea reală, a constatat că acest instrument AI nu a evidențiat probleme de siguranță și a îmbunătățit procesul de îngrijire, fără să demonstreze un beneficiu asupra rezultatelor pacienților în două săptămâni — iar detectarea unui beneficiu mic ar necesita un studiu mult mai mare.

De ce contează

Dezbaterea despre AI medicală duce lipsă tocmai de acest tip de dovezi. Există mii de lucrări în care modelele rezolvă examene și egalează clinicienii pe cazuri atent pregătite. Există însă foarte puține studii mari, pragmatice și randomizate care măsoară dacă pacienții reali au rezultate mai bune. Acesta este unul dintre ele și ajunge la o concluzie mai puțin spectaculoasă: a trece un test nu este același lucru cu a ajuta pacientul.

Această diferență este esențială. Un instrument poate fi cu adevărat util clinicienilor — note mai clare, costuri mai mici pentru medicamente, o a doua verificare — și totuși să nu modifice un rezultat clinic obiectiv în două săptămâni. Ambele lucruri pot fi adevărate simultan, iar un sistem de sănătate matur trebuie să le evalueze împreună, nu să aleagă doar concluzia convenabilă.

Studiul reașază și standardul dovezilor într-un loc mai util. Dacă o companie afirmă că AI-ul său clinic îmbunătățește îngrijirea, dovada relevantă nu este poziția într-un clasament, ci un studiu ca acesta, bazat pe rezultate care contează pentru pacienți — și, ideal, unul mai mare, pentru că beneficiile modeste necesită eșantioane mari pentru a putea fi detectate.

Pe scurt

Un studiu pragmatic, randomizat pe clustere, desfășurat în 16 unități de îngrijire primară din Kenya, a testat un instrument de suport decizional cu AI generativă („AI Consult") adăugat fișei electronice folosite de clinical officers. Dintre 9.691 de pacienți, indicatorul compozit de eșec terapeutic la 14 zile, evaluat de experți, a fost 2,2% cu instrumentul față de 2,0% fără (odds ratio ajustat 0,77, IC 95% 0,55–1,08, P = 0,13) — fără diferență semnificativă. Instrumentul nu a evidențiat probleme de siguranță, a îmbunătățit documentația clinică în toate domeniile evaluate, nu a schimbat prescrierea, nu a modificat satisfacția pacienților și a fost asociat cu costuri ceva mai mici pentru antibiotice. Autorii concluzionează că nu au apărut probleme de siguranță în limitele studiului, dar că eșecul terapeutic nu a fost redus; orice beneficiu este probabil modest, iar detectarea unui efect de mărimea celui observat ar necesita un studiu mult mai mare, de ordinul a 100.000 de pacienți. Rezultatul nu arată nici că AI clinică îmbunătățește rezultatele pacienților, nici că este inutilă: arată că un instrument care a îmbunătățit procesul de îngrijire nu a produs un beneficiu clinic demonstrabil în două săptămâni, în această rețea urbană privată.

Verificare fără exagerări

Ce arată lucrarea: Într-un studiu randomizat desfășurat în lumea reală, adăugarea unui instrument de suport decizional bazat pe LLM la fișele de îngrijire primară nu a evidențiat probleme de siguranță, a îmbunătățit calitatea documentației clinice, nu a schimbat prescrierea și nu a redus semnificativ indicatorul compozit strict de eșec terapeutic la 14 zile.

Ce este plauzibil, dar nedemonstrat: Că instrumentul produce o mică reducere reală a eșecului terapeutic, prea mică pentru ca acest studiu să o detecteze; că economisește bani după ce toate costurile sunt numărate; că documentația mai bună se traduce în cele din urmă în îngrijire mai bună.

Ce nu arată: Că AI clinică îmbunătățește rezultatele pacienților; că este nesigură pentru evenimente rare; că înlocuiește sau depășește clinicienii; că aceste rezultate se transferă în medii rurale sau cu venituri mai mari; că economia de cost este stabilită.

Limite principale: Putere pentru un efect mai mare decât cel observat (rezultatele rare cer eșantioane foarte mari); o eroare de configurare a contaminat brațul control, iar obiceiurile dintr-o rețea comună estompează comparația, ambele împingând spre nicio diferență; o singură rețea urbană privată cu standarde deja ridicate; niciun cadru prespecificat de noninferioritate sau siguranță; orizont scurt de 14 zile.

Câtă încredere ar trebui să aibă un cititor general? Mare că instrumentul nu a evidențiat probleme de siguranță în acest studiu și a îmbunătățit documentația. Mare și că nu a îmbunătățit demonstrabil rezultatele pacienților la 14 zile. Mică pentru orice afirmație categorică potrivit căreia „funcționează" sau „eșuează" ca intervenție menită să aducă beneficii pacienților — întrebarea rămâne deschisă și necesită un studiu mult mai mare. Poziția potrivită este să vedem aici un rezultat atent, din lumea reală, despre un instrument care a îmbunătățit procesul de îngrijire fără să demonstreze un beneficiu clinic, nu un verdict că AI transformă — sau distruge — medicina primară.

Surse

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