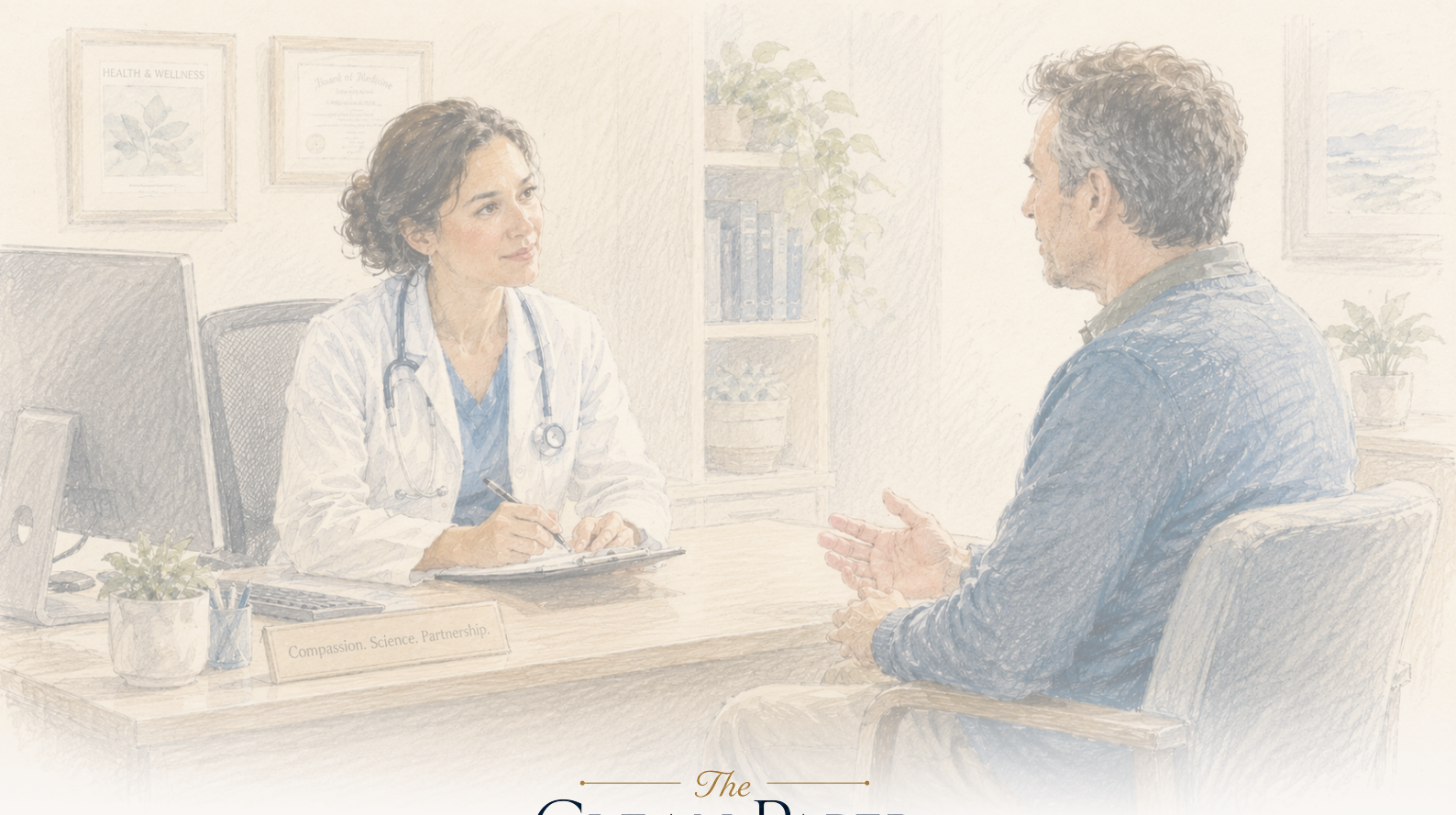


The Clean Paper

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

Një AI klinike që dukej e sigurt dhe përmirësoi dokumentimin — por jo rezultatet e pacientëve

Një provë pragmatike, e randomizuar në klasterë, në 16 qendra të kujdesit parësor në Kenia testoi një mjet mbështetës vendimmarrjeje me AI gjenerative të shtuar në kartelën e përdorur nga clinical officers. Mes 9,691 pacientëve, rezultati i rreptë i përbërë i dështimit të trajtimit brenda 14 ditësh, i gjykuar nga ekspertë, ishte 2.2% me mjetin kundrejt 2.0% pa të (raporti i rregulluar i odds 0.77, 95% CI 0.55–1.08, P = 0.13) — pa dallim domethënës. Mjeti nuk dha sinjal sigurie, përmirësoi dokumentimin në të gjitha fushat, la përshkrimin dhe kënaqësinë të pandryshuara dhe sinjalizoi kosto më të ulëta antibiotikësh. Ky nuk është studim i dështuar; është një studim i rrallë dhe i ndershëm — i matur te pacientët, jo benchmark-et — që tregon se një mjet që dukej i sigurt dhe ndihmonte procesin nuk i ndihmoi në mënyrë të demonstrueshme pacientët brenda dy javësh dhe se çdo përfitim i tillë do të kërkonte një provë shumë më të madhe.

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

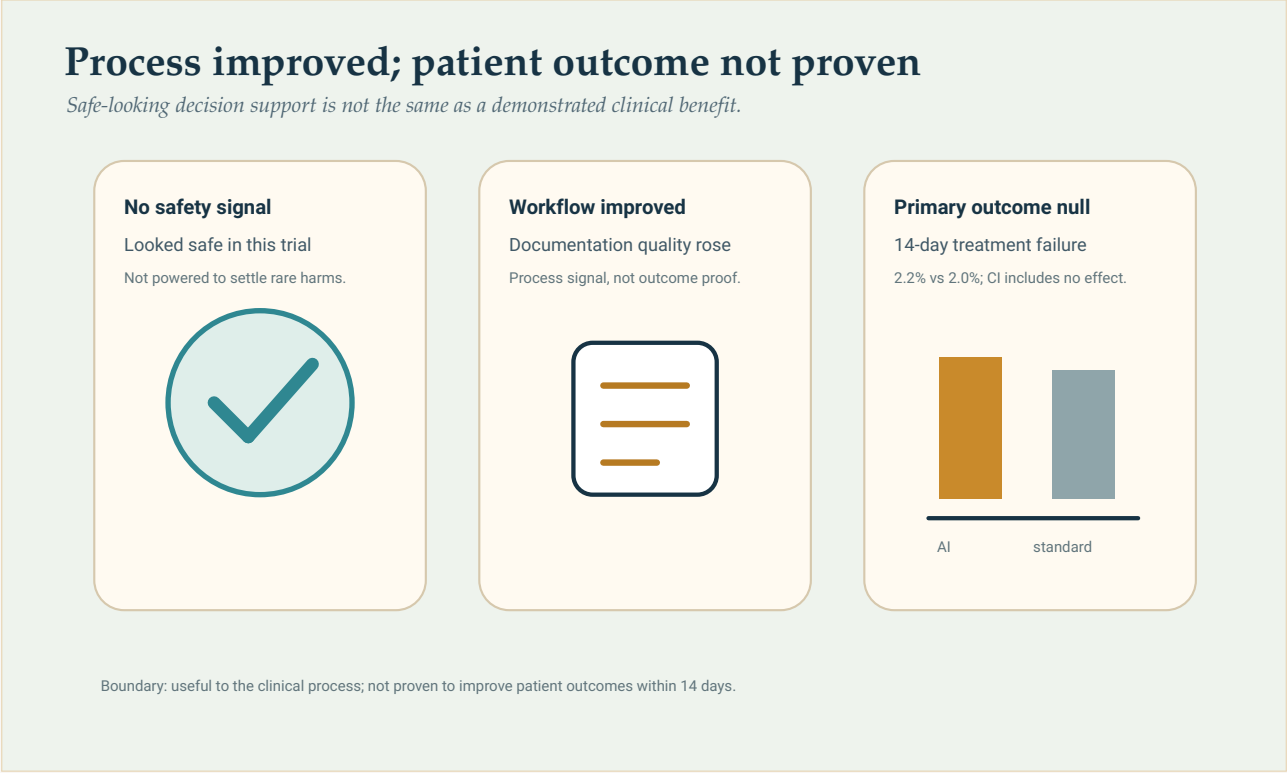
E sigurt dhe e dobishme nuk do të thotë domosdoshmërisht efektive

Shumica e titujve për AI mjekësore ndërtohen mbi llojin e gabuar të studimit. Një model del mirë në pyetje provimi, ose ia kalon mjekëve në raste klinike të përzgjedhura, dhe historia shkruhet vetë: makina është gati për klinikë.

Ky studim bëri diçka më të vështirë dhe shumë më të rrallë. Vendosi një mjet mbështetës vendim-marrjeje me AI gjenerative në kujdesin parësor real, në qendra reale, me pacientë realë, dhe pyeti atë që ka vërtet rëndësi: a dolën pacientët më mirë?

Përgjigjja e ndershme është jo — jo në mënyrë të matshme brenda 14 ditësh. Mjeti nuk dha sinjal sigurie. Përmirësoi cilësinë e dokumentimit klinik. Mund të ketë ulur edhe disa kosto barnash. Por nuk uli në mënyrë statistikisht domethënëse dështimet e trajtimit dhe autorët thonë me kujdes se çdo përfitim, nëse ekziston, ka gjasa të jetë modest.

Kjo nuk është dështim i studimit. Është studimi duke bërë punën e vet. Kështu duket prova e përgjegjshme për AI në mjekësi kur matet te pacientët dhe jo te benchmark-et.



Mjeti nuk dha sinjal sigurie dhe përmirësoi dokumentimin klinik, por rezultati i paracaktuar i pacientit në 14 ditë nuk u përmirësua në mënyrë statistikisht domethënëse. Ndihma në proces nuk është e njëjta gjë me përfitimin e provuar për pacientin.

Çfarë bënë autorët

Ekipi kreu një provë pragmatike, të randomizuar në klasterë, në **16 qendra të kujdesit parësor** të menaxhuara nga një rrjet privat shëndetësor (Penda Health) në qarqet Nairobi dhe Kiambu të Kenias. Kujdesi në këto qendra ofrohet kryesisht nga **clinical officers** — profesionistë shëndetësorë të nivelit të ndërmjetëm me diplomë trevjeçare — shpesh pa akses të lehtë në konsultë me mjekë më të vjetër.

Njësia e randomizimit ishte klinicisti, jo pacienti. **103 clinical officers** u randomizuan: 52 në grupin e ndërhyrjes dhe 51 në grupin kontroll. Të dy grupet përdornin të njëjtin kartelë elektronike shëndetësore në cloud. Grupi i ndërhyrjes kishte gjithashtu **AI Consult** (versioni 2.0), një mjet mbështetës vendimmarrjeje i ndërtuar mbi modelin e madh gjuhësor **GPT-4o të OpenAI-t** dhe i integruar në kartelë. Ai lexonte informacionin që dokumentonte klinicisti dhe mund të sinjalizonte probleme të mundshme me diagnozën ose planin e trajtimit. Klinicistët ruanin autonominë e plotë: mund t'i pranonin, ndryshonin ose injoronin sugjerimet.

> **Cili model ishte dhe pse detajet kanë rëndësi**
>
> Mjeti ishte **AI Consult 2.0**, me **GPT-4o të OpenAI-t** (versioni i majit 2025), i aksesuar përmes API-së komerciale të OpenAI-t nën licencë enterprise dhe i përdorur me parametra me pak rastësi (temperature 0.1). Ai ishte i integruar në një kartelë elektronike të ndërtuar posaçërisht (EMR-ja EasyClinic) dhe drejtohej nga system prompts të shkruara për t'u përputhur me udhëzimet kombëtare të trajtimit në Kenia; autorët publikuan prompt-in e plotë.
> Pse t'i themi këto? Sepse rezultati vlen për **një sistem specifik** — një version modeli, një prompt, një kartelë, një konfigurim — jo për "LLM-të në mjekësi" në përgjithësi. Autorët bëjnë të njëjtën vërejtje: e quajnë gjetjen e tyre një **benchmark kohor dhe jo një vlerësim fiks të aftësisë**. Një model më i ri, një prompt tjetër ose një klinikë më pak e dixhitalizuar mund të japë rezultat tjetër.
> Për pavarësinë: OpenAI më vonë ofroi mbështetje në natyrë (kredite cloud-compute dhe udhëzim teknik për përdorimin e API-së), por autorët deklarojnë se vendimi për të përdorur OpenAI ishte marrë **para** kësaj oferte dhe se OpenAI nuk kishte rol në dizajnin e provës, mbledhjen e të dhënave, analizën ose vendimin për publikim.

Mes 22 prillit dhe 16 korrikut 2025 u përfshinë **9,691 pacientë**. **Rezultati parësor ishte qëllimisht i përqendruar te pacienti dhe i rreptë**: një rezultat i përbërë, i gjykuar nga ekspertë, i **dështimit të trajtimit brenda 14 ditësh** nga vizita — një panel klinicistësh, i verbuar ndaj grupit të studimit, gjykoi nëse secili pacient kishte rezultat të keq si sëmundje e pazgjidhur ose e përkeqësuar. Prova ishte regjistruar paraprakisht (Pan-African Clinical Trials Registry 202502499779176).

Kjo zgjedhje dizajni është thelbi i punimit. Është e lehtë të tregosh se një mjet AI ndryshon çfarë shkruan klinicisti. Është shumë më e vështirë — dhe më kuptimplotë — të tregosh se ndryshon çfarë i ndodh pacientit.

Çfarë gjetën

Rezultati parësor nuk u përmirësua. Dështimi i trajtimit ndodhi te 102 nga 4,693 pacientë (2.2%) në grupin me AI dhe te 94 nga 4,654 (2.0%) në grupin kontroll. Përqindjet bruto ishin pak më të larta me AI, por pas rregullimit vlerësimi pikësor anon nga përfitimi: **raporti i rregulluar i odds ishte 0.77 (95% interval besimi 0.55 deri 1.08, P = 0.13)** — jo statistikisht domethënës. Ky ndryshim drejtimi mes numrave bruto dhe atyre të rregulluar nuk është gabim aritmetik; rregullimi llogarit dallimet mes klasterëve të klinicistëve. Sidoqoftë, intervali i besimit përfshin qartazi "asnjë efekt", prandaj përfitim nuk mund të pretendohet, dhe në terma absolutë efekti ishte shumë i vogël.

Për një mënyrë të thjeshtë për të lexuar së bashku odds ratios, intervalet e besimit dhe vlerat P, shih udhëzuesin për rezultatet klinike.

Asnjë sinjal sigurie — brenda kufijve të studimit. Asnjë ngjarje serioze e padëshiruar nuk u gjykua e lidhur me mjetin dhe një rishikim i pavarur nuk gjeti sinjal sigurie. Autorët janë të qartë për kufirin e këtij qetësimi: prova nuk kishte fuqi statistikore për të kapur dëme të rralla e të rënda dhe nuk kishte kornizë të paracaktuar non-inferiority apo sigurie formale, kështu që nuk mund të **provojë** siguri për ngjarje të rralla.

Dokumentimi u përmirësua. Në 2,000 konsultime të shqyrtuara nga ekspertë të verbuar, klinicistët që përdornin AI Consult prodhuan dokumentim më të mirë klinik në të gjitha fushat e vlerësuara — diagnozën e regjistruar, planin e trajtimit dhe plotësinë e përgjithshme.

Përshkrimi i barnave ndryshoi fare pak. Nuk kishte ndryshim domethënës në përshkrim, përfshirë përdorimin korrekt të antibiotikëve (raporti i rregulluar i odds 0.86, 95% CI 0.48 deri 1.55). Mjeti nuk ndryshoi normën e përshkrimit të antibiotikëve.

Pacientët nuk vunë re dallim. Mes 826 pacientëve që plotësuan një anketë kënaqësie, rezultatet ishin pothuajse identike mes dy grupeve dhe kohëzgjatja e konsultave ishte e ngjashme.

Kostot sinjalizuan një ulje të vogël. Në analizë të rregulluar, kostot e lidhura me antibiotikët ishin më të ulëta në grupin me AI — me gjasë përmes zgjedhjeve më të lira, jo përshkrimeve më të pakta — dhe kursimi për pacient dukej se tejkalonte koston për pacient të përdorimit të mjetit. Autorët e paraqesin këtë si sugjerim, jo rezultat të mbyllur: një llogaritje e plotë e kostos totale të pronësisë ishte jashtë fushës së provës.

Përmbledhja një-rreshtëshe e vetë autorëve është versioni më i pastër: ndihma me LLM dukej e sigurt brenda këtyre kufijve, por nuk uli dështimin e trajtimit brenda 14 ditësh, dhe çdo përfitim ka gjasa të jetë modest.

Pse “nuk ka dallim statistikisht domethënës” nuk do të thotë “nuk funksionon”

Një rezultat parësor nul mund të keqinterpretohet në të dy drejtimet. Dy gjëra pengojnë historinë e thjeshtë.

Së pari, **prova ishte ndërtuar për të kapur një efekt më të madh nga ai që u pa.** Rezultatet serioze të këqija në kujdesin parësor janë të rralla — rreth 2% këtu — kështu që për të dalluar një përfitim real të vogël nga zhurma duhen shumë më tepër pacientë. Llogaritjet post-hoc të autorëve sugjerojnë se për të detektuar një efekt të madhësisë së vëzhguar do të duhej një **provë shumë më e madhe, në rendin e mbi 100,000 pacientëve**. Një rezultat jo domethënës në një studim të kësaj madhësie nuk përjashton një përfitim të vogël real; do të thotë se ky studim nuk mund ta zgjidhte.

Së dyti, **krahasimi u turbullua pjesërisht.** Një gabim konfigurimi u dha për pak kohë disa klinicistëve në grupin kontroll akses në AI Consult dhe klinicistët në një rrjet të përbashkët flasin mes tyre dhe bartin zakone nga njëri grup në tjetrin. Të dy efektet priren t’i bëjnë grupet më të ngjashme dhe mund ta shtyjnë një dallim real drejt zeros. Për më tepër, rrjeti pritës tashmë punonte me standarde relativisht të larta, duke lënë më pak hapësirë për përmirësim.

Asnjëra prej këtyre nuk shpëton një titull “revolucion”. Por do të thotë se leximi i saktë duhet të jetë i kalibruar, jo përçmues: në rezultatin më të vështirë dhe më të ndershëm, ky mjet nuk tregoi përfitim të demonstrueshëm për pacientët brenda dy javësh — ndërsa përmirësoi qartë dokumentimin dhe dukej i sigurt.

Çfarë *nuk* provon kjo

- **Nuk tregon se AI përmirësoi rezultatet e pacientëve.** Në rezultatin parësor 14-ditor nuk kishte përfitim domethënës.
- **Nuk tregon se AI është e padobishme.** Vlerësimi pikësor anonte nga përfitimi, dokumentimi u përmirësua në të gjitha fushat dhe kostot e barnave sinjalizuan ulje; rezultati nul përputhet me një përfitim të vogël real që prova ishte shumë e vogël për ta konfirmuar.
- **Nuk provon se mjeti është i sigurt për dëme të rralla.** Nuk pati sinjal sigurie, por prova nuk kishte fuqi dhe dizajn për të certifikuar ngjarje të rralla e të rënda.
- **Nuk tregon se “AI ua kalon mjekëve” ose i zëvendëson klinicistët.** Kjo është **mbështetje vendim-marrjeje**; klinicisti ruante autoritetin e plotë për t’i pranuar ose refuzuar sugjerimet.
- **Nuk përgjithësohet automatikisht.** Prova u zhvillua në një rrjet të vetëm urban privat në Kenia; mjedise rurale, periurbane ose me të ardhura më të larta mund të japin rezultate të ndryshme në të dy drejtimet.
- **Nuk përcakton kursim kostosh.** Sinjali ekonomik është sugjerues, jo vlerësim i plotë ekonomik.

Sa e fortë është prova?

Për pretendimin qendror — se nuk u provua ulje e dëmit afatshkurtër te pacientët — prova është e fortë **si dizajn** dhe me të drejtë modeste **si përfundim**. Një provë prospektive, e regjistruar para-prakisht, e randomizuar në klasterë, me rezultat të përbërë në nivel pacienti dhe gjykim të verbuar, është afër llojit më të mirë të provës reale që mund të mblidhet për një mjet të tillë. Është shumë më informuese se një rezultat benchmark-u ose studim me vignette.

Për gjetjet dytësore — dokumentim më të mirë, përshkrim të pandryshuar, kënaqësi të ngjashme, kosto më të ulëta antibiotikësh — provat janë të mira, por duhen lexuar si dytësore: sinjale mbështetëse, jo titulli kryesor, dhe të ekspozuara ndaj të njëjtave kufizime të kontaminimit dhe rrjetit të vetëm.

Për sigurinë, prova është qetësuese por e kufizuar: nuk u gjet sinjal, por studimi nuk ishte ndërtuar për të kapur dëme të rralla.

Qëndrimi më i dobishëm nuk është as “funksionon”, as “dështoi”. Është: një provë reale dhe e kujdesshme gjeti se ky mjet AI nuk dha sinjal sigurie dhe përmirësoi procesin e kujdesit, pa demonstruar përfitim për rezultatin e pacientit brenda dy javësh — dhe çdo përfitim i tillë do të kërkonte një studim shumë më të madh për t’u parë.

Pse ka rëndësi

Debati për AI mjekësore vuan pikërisht nga mungesa e këtij lloj prove. Ka mijëra punime ku modele kalojnë provime dhe barazohen me klinicistët në raste të pastruara. Ka shumë pak prova të mëdha, pragmatike dhe të randomizuara që matin nëse pacientët realë përfitojnë. Ky është një prej tyre dhe mbërrin te e vërteta jo shumë glamuroze: të kalosh testin nuk është e njëjta gjë me të ndihmosh pacientin.

Ky hendek është e gjithë historia. Një mjet mund të jetë vërtet i dobishëm për klinicistët — shënime më të qarta, fatura barnash më të ulëta, një palë sy shtesë — dhe përsëri të mos lëvizë një rezultat të fortë pacienti brenda dy javësh. Të dyja mund të jenë të vërteta njëkohësisht dhe një sistem shëndetësor i pjekur duhet t’i mbajë bashkë, jo të zgjedhë vetëm njërën.

Rezultati gjithashtu zhvendos në një drejtim të dobishëm barrën e provës. Nëse një kompani dëshiron të pretendojë se AI e saj klinike përmirëson kujdesin, prova e rëndësishme nuk është një leaderboard. Është një provë si kjo, mbi rezultate që kanë rëndësi për pacientët — dhe, idealisht, një provë më e madhe, sepse mësimi i ndershëm këtu është se përfitimet modeste kërkojnë studime të mëdha për t’u parë.

Përmbledhje e shkurtër

Një provë pragmatike, e randomizuar në klasterë, në 16 qendra të kujdesit parësor në Kenia testoi një mjet mbështetës vendimmarrjeje me AI gjenerative ("AI Consult") të shtuar në kartelën elektronike që përdornin clinical officers. Mes 9,691 pacientëve, rezultati i përbërë i dështimit të trajtimit brenda 14 ditësh, i gjykuar nga ekspertë, ishte 2.2% me mjetin kundrejt 2.0% pa të (raporti i rregulluar i odds 0.77, 95% CI 0.55–1.08, P = 0.13) — pa dallim domethënës. Mjeti nuk dha sinjal sigurie, përmirësoi dokumentimin klinik në të gjitha fushat e vlerësuara, nuk ndryshoi përshkrimin e barnave, nuk ndryshoi kënaqësinë e pacientëve dhe u shoqërua me kosto disi më të ulëta të antibiotikëve. Autorët përfundojnë se dukej i sigurt brenda këtyre kufijve, por nuk uli dështimin e trajtimit, dhe çdo përfitim ka gjasa të jetë modest; për të detektuar një efekt të madhësisë së vëzhguar do të duhej një studim shumë më i madh, në rendin e 100,000 pacientëve. Rezultati nuk tregon as se AI klinike përmirëson rezultatet e pacientëve, as se është e padobishme — tregon se një mjet që dukej i sigurt dhe ndihmonte procesin nuk i ndihmoi në mënyrë të demonstrueshme pacientët brenda dy javësh në një rrjet të vetëm privat urban.

Kontroll pa teprime

Çfarë tregon punimi: Në një provë të randomizuar në botën reale, shtimi i një mjeti mbështetës me LLM në kartelat e kujdesit parësor nuk dha sinjal sigurie, përmirësoi cilësinë e dokumentimit klinik, nuk ndryshoi përshkrimin dhe nuk uli në mënyrë statistikisht domethënëse një rezultat të rreptë 14-ditor të përbërë të dështimit të trajtimit.

Çfarë është e besueshme, por jo e provuar: Që mjeti jep një ulje të vogël reale të dështimit të trajtimit, tepër të vogël për t'u kapur nga kjo provë; që kursen para kur llogariten të gjitha kostot; që dokumentimi më i mirë përkthehet më vonë në kujdes më të mirë.

Çfarë nuk tregon: Që AI klinike përmirëson rezultatet e pacientëve; që është e sigurt për ngjarje të rralla; që zëvendëson ose ua kalon klinicistëve; që rezultatet vlejnë njësoj në zona rurale ose vende me të ardhura më të larta; që kursimi i kostove është i vendosur.

Kufizimet kryesore: Fuqi statistikore për një efekt më të madh se ai i vëzhguar (rezultatet e rralla kërkojnë mostra shumë të mëdha); një gabim konfigurimi kontaminoi grupin kontroll dhe zakonet në rrjet të përbashkët zbehin krahasimin, të dyja duke e shtyrë dallimin drejt zeros; një rrjet privat urban i vetëm me standarde tashmë të larta; pa kornizë të paracaktuar non-inferiority ose sigurie; horizont i shkurtër 14-ditor.

Sa besim duhet të ketë një lexues i përgjithshëm? Të lartë se mjeti nuk dha sinjal sigurie në këtë provë dhe përmirësoi dokumentimin. Të lartë se nuk përmirësoi në mënyrë të demonstrueshme rezultatet 14-ditore të pacientëve këtu. Të ulët për çdo deklaratë se "funksionon" ose "dështoi" si ndërhyrje për përfitim pacienti — kjo pyetje mbetet realisht e pazgjidhur dhe kërkon një studim shumë më të madh. Qëndrimi i arsyeshëm: rezultat i kujdesshëm nga bota reale për një mjet që nuk dha sinjal sigurie dhe përmirësoi procesin e kujdesit, jo verdikt se AI transformon — ose shkatërron — kujdesin parësor.

Burimet

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