

The Clean Paper

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

Клиничка AI што изгледала безбедно и ја подобрила документацијата — но не ги подобрила исходите за пациентите

Прагматично, cluster-randomized испитување во 16 установи за примарна здравствена заштита во Кенија тестирало генеративна AI-алатка за поддршка при одлучување. Кај 9.691 пациенти, строгиот 14-дневен составен исход на неуспех на лекувањето бил 2,2% со алатката наспроти 2,0% без неа (adjusted odds ratio 0,77, 95% CI 0,55–1,08, P = 0,13) — без значајна разлика. Алатката не покажала безбедносен сигнал, ја подобрила документацијата во сите оценувани домени, не го променила препишувањето или задоволството и покажала сигнал за пониски трошоци за антибиотици. Тоа не е неуспешна студија; тоа е ретка, искрена студија измерена врз пациенти, а не benchmark-и — и покажува дека алатка што изгледала безбедно и помогнала во процесот не им помогнала мерливо на пациентите во две недели, а за откривање евентуална умерена корист би била потребна многу поголема студија.

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

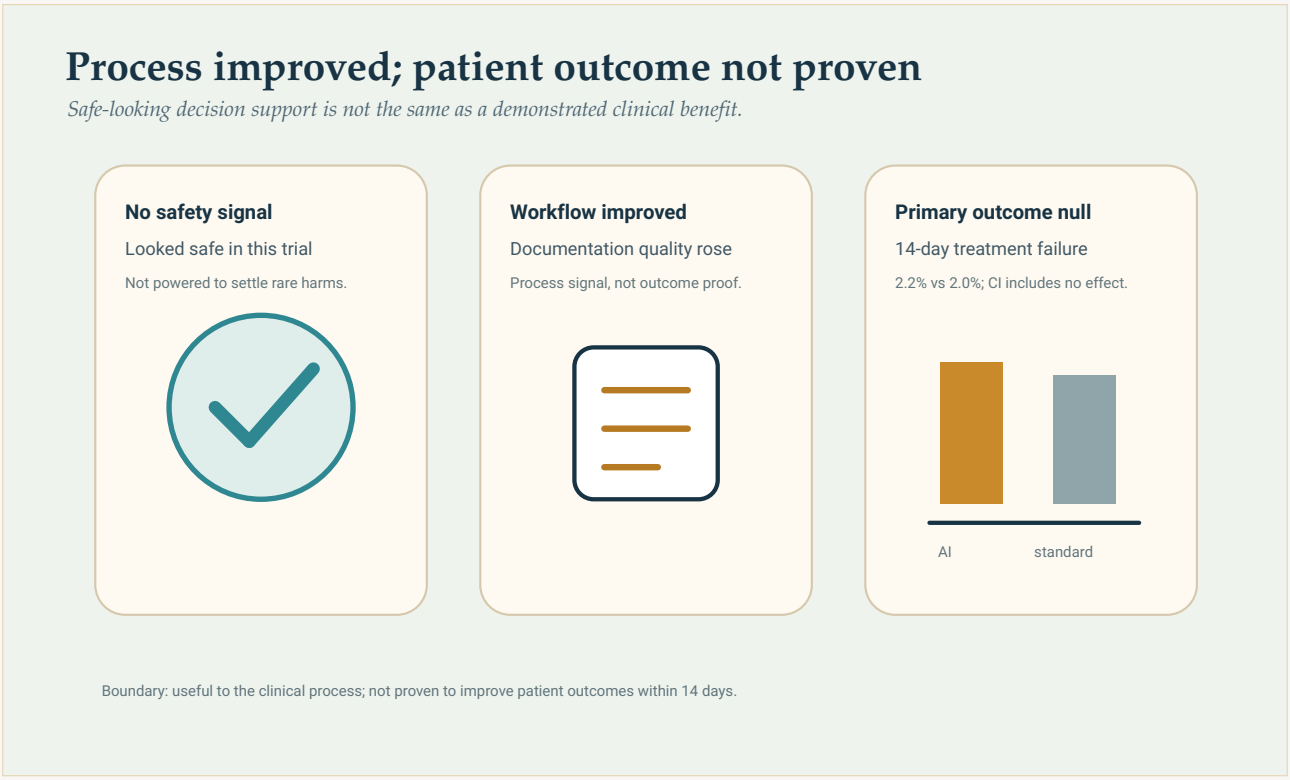
Безбедно и корисно не е исто што и ефикасно

Повеќето наслови за медицинска AI се градат врз погрешен тип студија. Моделот постигнува добар резултат на испитни прашања или ги надминува лекарите на внимателно избрани клинички сценарија, и приказната сама се пишува: машината е подготвена за клиника.

Ова испитување направи нешто потешко и поретко. Алатка за поддршка при одлучување со генеративна AI беше ставена во вистинска примарна здравствена заштита, во вистински установи и со вистински пациенти, а прашањето беше она што навистина е важно: дали пациентите поминале подобро?

Искреniот одговор е не — барем не мерливо во рок од 14 дена. Алатката не покажа безбедносен сигнал. Го подобри квалитетот на клиничката документација. Можно е дури и да намалила дел од трошоците за лекови. Но не го намалила значајно неуспехот на лекувањето, а авторите внимателно велат дека секоја корист, ако постои, веројатно е умерена.

Тоа не е неуспех на студијата. Тоа е студијата што си ја врши работата. Вака изгледа одговорно собирање докази за AI во медицината кога се мерат исходи кај пациенти, а не резултати на референтни тестови.



Алатката не покажа безбедносен сигнал и ја подобри клиничката документација, но однапред дефинираниот 14-дневен исход за пациентите не се подобри статистички значајно. Помош во процесот не е исто што и докажана корист за пациентот.

The Clean Paper CC BY 4.0

Што направиле авторите

Тимот спровел прагматично, кластерски рандомизирано испитување во **16 установи за примарна здравствена заштита** на приватната мрежа Penda Health во округзите Најроби и Киамбу, Кенија. Грижата во овие установи во голема мера ја обезбедуваат **clinical officers** — здравствени професионалци од средно ниво со тригодишна диплома — често без лесен пристап до постар лекар за консултација.

Единица на рандомизација бил клиничарот, не пациентот. **103 clinical officers** биле рандомизирани: 52 во интервенциската и 51 во контролната група. Двете групи го користеле истиот електронски медицински картон во облак. Интервенциската група дополнително имала **AI Consult** (верзија 2.0), алатка за поддршка при одлучување изградена врз големиот јазичен модел **GPT-4o на OpenAI** и вградена во самиот картон. Таа ги читала податоците што ги внесувал клиничарот и можела да сигнализира можни проблеми со дијагнозата или планот на лекување. Клиничарите ја задржале целата автономија: можеле да ги прифатат, изменат или игнорираат предлозите.

> **Кој модел бил користен и зошто деталите се важни**
>
> Алатката била **AI Consult 2.0**, со **GPT-4o на OpenAI** (изданието од мај 2025), достапуван преку комерцијалниот API на OpenAI со enterprise лиценца и со ниска случајност во генерирањето (temperature 0.1). Била вградена во посебно дизајниран електронски картон, EasyClinic EMR, а системските инструкции биле напишани така што да се усогласат со кениските национални водичи за лекување; авторите го објавиле целиот instruction prompt.
> Зошто да се наведува сето ова? Бидејќи резултатот се однесува на *еден конкретен систем* — една верзија на модел, еден prompt, еден електронски картон и една конфигурација — а не на „LLM-и во медицината“ воопшто. И авторите го нагласуваат истото: својот резултат го нарекуваат *временска референтна точка, а не фиксна процена на способност*. Понов модел, поинаков prompt или клиника со помала дигитализација може да дадат поинаков исход.
> За независноста: OpenAI подоцна обезбедил in-kind поддршка, со cloud-compute кредити и технички насоки за API, но авторите велат дека одлуката да се користи OpenAI била донесена *пред* таа понуда и дека OpenAI немал улога во дизајнот на испитувањето, собирањето или анализата на податоците ниту во одлуката за објавување.

Меѓу 22 април и 16 јули 2025 биле вклучени **9.691 пациенти**. **Примарниот исход бил намерно строг и насочен кон пациентот**: составен исход на *неуспех на лекувањето* во рок од 14 дена од посетата, проценет од експертска комисија. Панел клиничари, без да знае во која група бил пациентот, проценувал дали се случил неповолен исход, како нерешена или влошена болест. Испитувањето било однапред регистрирано во Pan-African Clinical Trials Registry (202502499779176).

Токму тој избор на дизајн е поентата. Лесно е да се покаже дека AI-алатка го менува она

што клиничарот го запишува. Многу потешко — и многу позначајно — е да се покаже дека го менува она што му се случува на пациентот.

Што откриле

Примарниот исход не се подобрил. Неуспех на лекувањето се случил кај 102 од 4.693 пациенти (2,2%) во AI-групата и 94 од 4.654 (2,0%) во контролната група. Суровите проценти биле малку повисоки со AI, но по статистичко приспособување точковната процена се наклонила кон корист: **приспособениот odds ratio бил 0,77 (95% confidence interval 0,55 до 1,08, P = 0,13)** — без статистичка значајност. Промената на насоката меѓу суровите и приспособените бројки не е аритметичка грешка; приспособувањето ги зема предвид разликите меѓу кластерите на клиничари. Во секој случај, интервалот на доверба јасно го вклучува „нема ефект“, па корист не може да се тврди, а апсолутниот ефект е мал.

За едноставно објаснување како заедно се читаат odds ratio, confidence interval и P-вредност, видете го водичот за читање клинички резултат.

Нема безбедносен сигнал — во границите на испитувањето. Ниеден сериозен несакан настан не бил оценет како поврзан со алатката, а независен преглед не нашол безбедносен сигнал. Авторите искрено ја наведуваат границата на ова уверување: испитувањето немало доволна статистичка моќ за ретки тешки штети и немало однапред дефинирана рамка за неинфериорност или формална безбедносна рамка, па не може да *докаже* безбедност за невообичаени настани.

Документацијата станала подобра. Меѓу 2.000 средби прегледани од експерти што не знаеле во која група е случајот, клиничарите со AI Consult имале подобра клиничка документација во сите оценувани домени — запишаната дијагноза, планот на лекување и општата целосност.

Препишувањето речиси не се поместило. Немало значајна разлика во препишувањето, вклучително и правилната употреба на антибиотици (adjusted odds ratio 0,86, 95% CI 0,48 до 1,55). Алатката не ја променила стапката на препишување антибиотици.

Пациентите не забележале разлика. Кај 826 пациенти што пополниле анкета, задоволството било практично исто во двете групи, а времетраењето на консултациите било слично.

Трошоците благо се насочиле надолу. Во приспособена анализа, трошоците поврзани со антибиотици биле пониски во AI-групата — веројатно преку избор на поевтини лекови, не преку помалку рецепти — и заштедата по пациент изгледала поголема од трошокот за работа на алатката по пациент. Авторите го означуваат ова како сигнал, не како утврден економски резултат: целосна анализа на total cost of ownership не била дел од испитувањето.

Најчистото едноречениично резиме на авторите е: LLM-поддршката била безбедна во

границите на студијата, но не го намалила неуспехот на лекувањето во 14 дена, а секоја корист веројатно е умерена.

Зошто „нема значајна разлика“ не значи „не работи“

Нултиот примарен исход лесно може да се претолкува во двете насоки. Две работи ја спречуваат едноставната приказна.

Прво, **испитувањето било дизајнирано да открие поголем ефект од оној што го забележало**. Сериозните неповолни исходи во примарна здравствена заштита се ретки — околу 2% тука — па разликување мал реален ефект од шум бара огромен број пациенти. Post-hoc пресметките на авторите сугерираат дека за ефект со големина каква што набљудувале би бил потребен **многу поголем trial, од редот на повеќе од 100.000 пациенти**. Незначаен резултат во студија со оваа големина не исклучува мал, реален ефект; значи дека оваа студија не можела да го разреши.

Второ, **споредбата делумно била заматена**. Конфигурациска грешка накратко им дала пристап до AI Consult на некои клиничари од контролната група, а клиничари во иста мрежа разговараат и пренесуваат навики преку границите на групите. И двете работи ги прават групите послични и можат вистинската разлика да ја турнат кон нула. Дополнително, мрежата веќе работела со релативно високи стандарди, оставајќи помал простор за подобрување.

Ништо од ова не спасува наслов „пробив“. Но значи дека точниот прочит е калибриран, не дефлациски: на најтешкиот и најискрен исход, оваа алатка не покажала мерлива корист за пациентите во две недели — додека јасно ја подобрила документацијата и изгледала безбедно.

Што ова не докажува

- Не покажува дека AI ги подобрила исходите за пациентите. На примарниот 14-дневен исход немало значајна корист.
- Не покажува дека AI е бескорисна. Точковната процена била во корист на алатката, документацијата се подобрила, а трошоците за лекови се насочиле надолу; нултиот резултат е согласен и со мал реален ефект што студијата била премала да го потврди.
- Не докажува дека алатката е безбедна за ретки штети. Нема пронајден безбедносен сигнал, но студијата не била дизајнирана да сертифицира ретки сериозни настани.
- Не покажува дека „AI ги победува лекарите“ или ги заменува клиничарите. Ова е поддршка при одлучување; клиничарот ја задржал целата власт да прифати или одбие предлог.
- Не се генерализира автоматски. Испитувањето е во една приватна урбана мрежа во Кенија; рурални, периурбани и побогати средини можат да се разликуваат во која

било насока.
- Не утврдува заштеда на трошоци. Економскиот сигнал е сугестивен, а не завршена економска евалуација.

Колку се убедливи доказите?

За централното тврдење — нема докажано намалување на краткорочниот неуспех на лекувањето — доказите се силни *по дизајн* и соодветно скромни *по заклучок*. Проспективно, однапред регистрирано, кластерски рандомизирано испитување со заслепено оценување и составен исход на ниво на пациент е блиску до најдобриот доказ од реалната клиничка практика што може да се собере за ваква алатка. Тоа е многу поинформативно од резултат на референтен тест или студија со клинички сценарија.

За секундарните резултати — подобра документација, непроменето препишување, слично задоволство, пониски трошоци за антибиотици — доказите се добри, но треба да се читаат како секундарни: поддржувачки сигнали, не главниот резултат, и подложни на истите ограничувања од контаминацијата и единствената мрежа.

За безбедноста, доказите се охрабрувачки, но ограничени: нема сигнал, но ова не е студија направена да открива ретки штети.

Најкорисниот став не е ниту „работи“, ниту „не успеа“. Туку: внимателно реално испитување откри дека оваа AI-алатка не покажала безбедносен сигнал и го подобрила процесот на грижа, без да покаже корист врз исходот за пациентот во две недели — а откривањето таква корист би барало многу поголема студија.

Зошто е важно

Дебатата за медицинска AI е гладна токму за ваков доказ. Постојат илјадници трудови во кои модели решаваат испити и се споредуваат со клиничари на уредни случаи. Многу малку се големи, прагматични, рандомизирани испитувања што мерат дали вистинските пациенти поминуваат подобро. Ова е едно од нив, и завршува во неатрактивната вистина: да го положиш тестот не е исто што и да му помогнеш на пациентот.

Токму таа празнина е целата приказна. Алатка може навистина да им биде корисна на клиничарите — појасни белешки, пониски сметки за лекови, втор пар очи — и сепак да не помести тврд пациентски исход за две недели. Двете работи можат истовремено да бидат вистина, а зрел здравствен систем мора да ги држи заедно наместо да избере само една удобна верзија.

Студијата корисно го поставува и товарот на доказ. Ако компанија сака да тврди дека

нејзината клиничка AI ја подобрува грижата, релевантниот доказ не е leaderboard. Тоа е испитување како ова, на исходи што им се важни на пациентите — и идеално поголемо, бидејќи искрената лекција тука е дека за умерени придобивки се потребни големи студии.

Кратко резиме

Прагматично, кластерски рандомизирано испитување во 16 установи за примарна здравствена заштита во Кенија тестирало генеративна AI-алатка за поддршка при одлучување, AI Consult, додадена во електронскиот картон на clinical officers. Кај 9.691 пациенти, експертски проценетиот составен исход на неуспех на лекувањето во 14 дена бил 2,2% со алатката наспроти 2,0% без неа (прилагоден однос на шансите 0,77, 95% интервал на доверба 0,55–1,08, P = 0,13) — без значајна разлика. Алатката не покажала безбедносен сигнал, ја подобрила клиничката документација во сите оценувани домени, не го променила препишувањето, не го променила задоволството на пациентите и била поврзана со малку пониски трошоци за антибиотици. Авторите заклучуваат дека била безбедна во тие граници, но не го намалила неуспехот на лекувањето; секоја корист веројатно е умерена, а за ефект со набљудуваната големина би била потребна многу поголема студија, од редот на 100.000 пациенти. Резултатот не покажува дека клиничката AI ги подобрува исходите, ниту дека е бескорисна — покажува дека алатка што изгледала безбедно и го подобрила процесот не покажала мерлива корист за пациентите во две недели, во една урбана приватна мрежа.

Проверка без претерување

Што покажува трудот: во реално рандомизирано испитување, додавањето LLM-алатка за поддршка во примарната грижа не покажало безбедносен сигнал, ја подобрило клиничката документација, не го променило препишувањето и не го намалило значајно строгиот 14-дневен составен исход на неуспех на лекувањето.

Што е веројатно, но не е докажано: дека алатката создава мал реален пад на неуспехот што оваа студија била премала да го открие; дека штеди пари кога ќе се пресметаат сите трошоци; дека подобрата документација подоцна води кон подобра грижа.

Што не покажува: дека клиничката AI ги подобрува исходите; дека е безбедна за ретки настани; дека ги заменува или надминува клиничарите; дека резултатите автоматски важат во рурални или побогати средини; дека економската заштеда е утврдена.

Главни ограничувања: студијата била подготвена за поголем ефект од забележаниот, а ретките исходи бараат огромни примероци; конфигурациска грешка ја контаминирала контролната група и навиките во заедничка мрежа ја заматуваат споредбата, што

може да ја намали разликата; една приватна урбана мрежа со веќе високи стандарди; нема однапред дефинирана рамка за неинфериорност или безбедност; краток хоризонт од 14 дена.

Колку доверба треба да има општ читател? Висока дека алатката не покажала безбедносен сигнал во ова испитување и ја подобрила документацијата. Висока дека тука не покажала мерливо подобрување на 14-дневните исходи. Ниска за едноставно тврдење дека „работи“ или „не работи“ како интервенција за корист на пациентот — тоа прашање навистина останува нерешено и бара многу поголема студија. Соодветниот став е: внимателен реален резултат за алатка што изгледала безбедно и го подобрила процесот на грижа, не пресуда дека AI ја трансформира — или ја уништува — примарната здравствена заштита.

Извори

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