

The Clean Paper

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

Клиничен AI, който изглеждаше безопасен и подобри документацията — но не подобри изходите за пациентите

Прагматично, клъстер-рандомизирано изпитване в 16 центъра за първична помощ в Кения тества инструмент за клинична подкрепа с генеративен AI, добавен към електронното досие на clinical officers. При 9 691 пациенти строгият 14-дневен комбиниран показател за неуспех на лечението, оценен от експерти, е 2,2% с инструмента срещу 2,0% без него (коригирано odds ratio 0,77, 95% CI 0,55–1,08, P = 0,13) — без значима разлика. Инструментът не показа сигнал за безопасност, подобри документацията във всички оценявани области, не промени предписването или удовлетвореността и показа тенденция към по-ниски разходи за антибиотици. Това не е провалено изследване; това е рядко честно изследване — измерено върху пациенти, не benchmark-и — и стига до негламурната истина, че инструмент, който изглежда безопасен и помага на процеса, не демонстрира полза за пациентите за две седмици, а откриването на такава полза би изисквало много по-голямо изпитване.

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

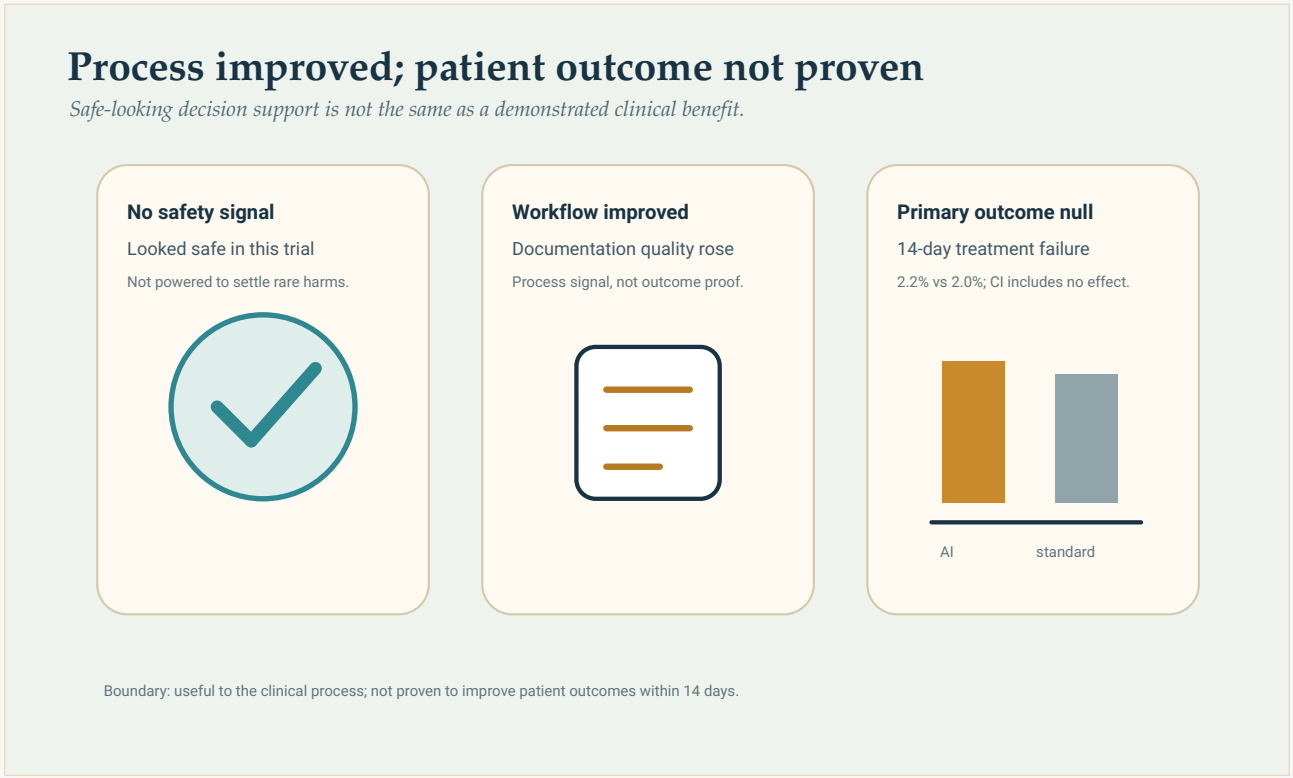
Безопасно и полезно не е същото като ефективно

Повечето заглавия за медицински AI са построени върху грешния вид изследване. Моделът получава висок резултат на изпитни въпроси или превъзхожда лекари върху подбрани клинични винетки и историята се пише сама: машината е готова за клиниката.

Това изпитване прави нещо по-трудно и по-рядко. Поставя генеративен AI за подпомагане на решенията в реална първична помощ, в реални лечебни центрове, с реални пациенти, и задава въпроса, който действително има значение: пациентите справят ли се по-добре?

Честният отговор е не — поне не измеримо и не за 14 дни. Инструментът не показва сигнал за безопасност. Подобрява качеството на клиничната документация. Може дори да е намалил някои разходи за лекарства. Но не намалява значимо неуспехите на лечението и авторите внимателно казват, че ако има полза, тя вероятно е умерена.

Това не е провал на изследването. Това е изследването, което работи. Така изглеждат отговорните доказателства за AI в медицината, когато се измерват върху пациенти, а не върху тестове за оценяване.



Инструментът не показа сигнал за безопасност и подобри клиничната документация, но предварително зададеният 14-дневен резултат за пациентите не се подобри значимо. Помощта за процеса не е същото като доказана полза за пациента.

The Clean Paper CC BY 4.0

Какво направиха авторите

Екипът провежда прагматично, клъстер-рандомизирано изпитване в **16 центъра за първична помощ**, управлявани от частна здравна мрежа (Penda Health) в окръзите Nairobi и Kiambu, Кения. Грижата в тези центрове се предоставя главно от **клинични специалисти (clinical officers)** — специалисти от междинно ниво с тригодишна диплома — често без лесен достъп до старши консултация.

Единицата на рандомизация е клиницистът, не пациентът. **103 клинични специалисти** са рандомизирани: 52 в интервенционната група и 51 в контролната. И двете групи използват една и съща облачна електронна медицинска система. Интервенционната група допълнително има **„AI Consult“** (версия 2.0) — инструмент за подкрепа на решенията, изграден върху големия езиков модел **GPT-4o на OpenAI** и вграден в досието. Той прочита информацията, документирана от клиницист, и може да сигнализира за възможни проблеми с диагнозата или лечебния план. Клиницистите запазват пълна автономия: могат да приемат, променят или игнорират предложенията му.

> **Кой модел е използван и защо детайлите имат значение**
>
> Инструментът е **AI Consult 2.0**, работещ с **GPT-4o на OpenAI** (изданието от май 2025 г.), достъпен през комерсиалния API на OpenAI с корпоративен лиценз и при ниска случайност (температура 0,1). Той е вграден в специално електронно досие (EasyClinic’s EMR) и е насочван чрез системни инструкции, написани така, че да съответстват на националните клинични насоки на Кения; авторите публикуват цялата инструкция.
> Защо да уточняваме всичко това? Защото резултатът е за *една конкретна система* — една версия на модел, една инструкция, едно досие, една среда — не за „LLM в медицината“ по принцип. Авторите правят същото уточнение: наричат находката си *моментна снимка във времето, а не фиксирана оценка на способностите*. По-нов модел, различна инструкция или по-слабо дигитализирана клиника могат да променят резултата.
> За независимостта: по-късно OpenAI предоставя помощ в натура (кредити за облачни изчисления и технически насоки за API), но авторите заявяват, че решението да използват OpenAI е взето *преди* това предложение и че OpenAI няма роля в дизайна на изпитването, събирането или анализа на данните или решението за публикация.

Между 22 април и 16 юли 2025 г. са включени **9 691 пациенти**. **Основният резултат е нарочно пациентски и строг**: експертно оценен *комбиниран показател за неуспех на лечението* в рамките на 14 дни след посещението — панел от клиницисти, без да знае групата, преценява дали всеки пациент е имал лош изход, например непреминало или влошено заболяване. Изпитването е регистрирано предварително (Pan-African Clinical Trials Registry 202502499779176).

Този избор на дизайн е самият смисъл. Лесно е да покажете, че AI инструмент променя какво записва клиницистът. Много по-трудно — и много по-значимо — е да покажете, че променя какво се случва с пациента.

Какво откриха

Основният резултат не се подобрява. Неуспех на лечението настъпва при 102 от 4 693 пациенти (2,2%) в AI групата и 94 от 4 654 (2,0%) в контролната. Суровите проценти са малко по-високи с AI, но след корекция точковата оценка се накланя към полза: **коригираното отношение на шансовете е 0,77 (95% доверителен интервал 0,55 до 1,08, P = 0,13)** — без статистическа значимост. Обръщането между суровите и коригираните числа не е аритметична грешка; корекцията отчита разлики между клъстерите от клиницисти. И в двата случая доверителният интервал спокойно включва „никакъв ефект“, така че полза не може да се заяви, а в абсолютни стойности ефектът е много малък.

За лесен начин да се четат отношение на шансовете, доверителни интервали и P стойности заедно, вижте ръководството за клинични резултати.

Няма сигнал за безопасност — в рамките на ограниченията. Нито едно сериозно нежелано събитие не е преценено като свързано с инструмента и независим преглед не открива сигнал за безопасност. Авторите честно посочват тавана на това успокоение: изпитването няма мощност да открива редки тежки вреди и няма предварително зададена рамка за доказване на не по-ниска ефективност или формална безопасност, така че не може да *докаже* безопасност за редки събития.

Документацията става по-добра. Сред 2 000 посещения, прегледани от заслепени експерти, клиницистите с AI Consult създават по-добра клинична документация във всички оценени области — записаната диагноза, лечебния план и общата пълнота.

Предписването почти не се променя. Няма значима разлика в предписването, включително в правилната употреба на антибиотици (коригирано отношение на шансовете 0,86, 95% CI 0,48 до 1,55). Инструментът не променя честотата на предписване на антибиотици.

Пациентите не забелязват разлика. Сред 826 пациенти, попълнили анкета за удовлетвореност, удовлетворението е практически еднакво в двете групи, а времетраенето на консултациите е подобно.

Разходите леко сочат надолу. В коригиран анализ разходите, свързани с антибиотици, са по-ниски в AI групата — вероятно чрез избор на по-евтини лекарства, не чрез по-малко рецепти — и спестеното на пациент изглежда надвишава разхода за работа на инструмента. Авторите го отбелязват като подсказка, не като установен резултат: пълна оценка на общата цена на притежание не е част от изпитването.

Едно изречение на самите автори е най-чистото резюме: LLM помощта е безопасна в

тези граници, но не намалява неуспеха на лечението за 14 дни, а всяка полза вероятно е умерена.

Защо „няма значима разлика“ не означава „не работи“

Нулевият основен резултат лесно може да бъде прочетен прекалено силно и в двете посоки. Две неща спират простата история.

Първо, **изпитването е проектирано да улови по-голям ефект от намерения**. Сериозните лоши изходи в първичната помощ са редки — тук около 2% — затова разграничаването на малка реална полза от шум изисква огромен брой хора. Собствените последващите (post-hoc) изчисления на мощността на авторите подсказват, че за ефект с размера на наблюдавания би било необходимо **много по-голямо изпитване, от порядъка на над 100 000 пациенти**. Незначим резултат в изследване с този размер не изключва малка реална полза; означава, че изследването не може да я разреши.

Второ, **сравнението е частично размито**. Конфигурационна грешка за кратко дава достъп до AI Consult на някои клиницисти от контролната група, а клиницисти в обща мрежа разговарят и пренасят навици през границата между групите. И двата ефекта правят групите по-подобни и бутат реална разлика към нула. Освен това мрежата домакин вече работи при сравнително високи стандарти, което оставя по-малко пространство за подобрение.

Нищо от това не спасява заглавие „пробив“. Но означава, че правилният прочит е калибриран, не унищожителен: на най-трудния и честен показател този инструмент не демонстрира полза за пациентите в две седмици — докато измеримо подобрява документацията и изглежда безопасен.

Какво *не* доказва това

- То **не** показва, че AI е подобрил резултатите за пациентите. На основния 14-дневен показател няма значима полза.
- То **не** показва, че AI е безполезен. Точковата оценка го предпочита, документацията се подобрява навсякъде, а разходите за лекарства се движат надолу; нулевият резултат е съвместим с малка реална полза, която изпитването е твърде малко да потвърди.
- То **не** доказва, че инструментът е безопасен за редки вреди. Не се открива сигнал за безопасност, но изпитването не е проектирано или достатъчно мощно да сертифицира безопасност за редки тежки събития.
- То **не** показва, че „AI побеждава лекарите“ или ги заменя. Това е *подкрепа* на решенията; клиницистът запазва пълното право да приеме или отхвърли предложението.
- То **не** се обобщава автоматично. Изпитването е в една частна градска мрежа в Кения;

селски, пери-градски и по-богати среди може да се различават в двете посоки.
- То **не** установява спестяване на разходи. Сигналът за разходите е подсказващ, не завършена икономическа оценка.

Колко силни са доказателствата?

За централното твърдение — няма доказано намаление на краткосрочната вреда за пациента — доказателствата са силни *като дизайн* и подходящо скромни *като заключение*. Проспективно, предварително регистрирано, клъстер-рандомизирано изпитване със заслепен комбиниран резултат на ниво пациент е близо до най-доброто реално доказателство, което може да се събере за такъв инструмент. То е много по-информативно от резултат от тест за оценяване или изследване с винетки.

За вторичните находки — по-добра документация, непроменено предписване, подобна удовлетвореност, по-ниски разходи за антибиотици — доказателствата са добри, но трябва да се четат като вторични: подкрепящи сигнали, не заглавието, и уязвими на същото замърсяване и ограничение до една мрежа.

За безопасността доказателствата са успокояващи, но ограничени: няма намерен сигнал, но това не е изследване, построено да открива редки вреди.

Най-полезната позиция не е нито „работи“, нито „провали се“. Тя е: внимателно реално изпитване установи, че този AI инструмент не показва сигнал за безопасност и подобрява процеса на грижа, без да демонстрира полза за пациента в рамките на две седмици — и че за откриване на такава полза би било необходимо много по-голямо изследване.

Защо това има значение

Дебатът за медицинския AI гладува точно за такива доказателства. Има хиляди статии, в които модели решават изпити и се изравняват с клиницисти по подредени случаи. Има много малко големи прагматични рандомизирани изпитвания, които измерват дали реалните пациенти са по-добре. Това е едно от тях и стига до негламурната истина: да издържиш теста не е същото като да помогнеш на пациента.

Тази пропаст е цялата история. Един инструмент може да е истински полезен за клиницистите — по-ясни бележки, по-ниски лекарствени сметки, втори чифт очи — и пак да не промени труден пациентски резултат за две седмици. И двете могат да са верни едновременно и зряла здравна система трябва да ги държи заедно, вместо да избира удобното.

Това пренареждa и тежестта на доказване в полезна посока. Ако компания иска да твърди, че клиничният ѝ AI подобрява грижата, релевантното доказателство не е

класация. То е изпитване като това, върху резултати, важни за пациентите — и в идеалния случай по-голямо, защото честният урок тук е, че умерените ползи изискват големи изследвания, за да бъдат видени.

Кратко и ясно

Прагматично, клъстер-рандомизирано изпитване в 16 центъра за първична помощ в Кения тества генеративен AI за подкрепа на решенията („AI Consult“), добавен към електронното досие, използвано от клинични специалисти. Сред 9 691 пациенти експертно оценениият комбиниран показател за неуспех на лечението в рамките на 14 дни е 2,2% с инструмента срещу 2,0% без него (коригирано отношение на шансовете 0,77, 95% CI 0,55–1,08, P = 0,13) — без значима разлика. Инструментът не показва сигнал за безопасност, подобрява клиничната документация във всички оценени области, не променя предписването, не променя удовлетвореността на пациентите и е свързан с малко по-ниски разходи за антибиотици. Авторите заключават, че е безопасен в тези граници, но не намалява неуспеха на лечението, като всяка полза вероятно е умерена; откриването на ефект с наблюдавания размер би изисквало много по-голямо изпитване (от порядъка на 100 000 пациенти). Резултатът не показва нито че клиничният AI подобрява изходите за пациентите, нито че е безполезен — показва, че инструмент, който изглежда безопасен и помага на процеса, не демонстрира полза за пациентите за две седмици в една частна градска мрежа.

Проверка без излишен шум

Какво показва статията: В реално рандомизирано изпитване добавянето на LLM-базиран инструмент за подкрепа към първичната помощ не показва сигнал за безопасност, подобрява качеството на клиничната документация, не променя предписването и не намалява значимо строг 14-дневен комбиниран показател за неуспех на лечението.

Какво е правдоподобно, но не е доказано: Че инструментът носи малко реално намаление на неуспеха на лечението, твърде малко за това изпитване; че спестява пари след пълно отчитане на всички разходи; че по-добрата документация в крайна сметка води до по-добра грижа.

Какво не показва: Че клиничният AI подобрява пациентските резултати; че е безопасен за редки събития; че заменя или превъзхожда клиницистите; че резултатите се пренасят към селски или по-богати среди; че спестяването на разходи е установено.

Основни ограничения: Изпитването е оразмерено за по-голям ефект от наблюдавания (редките резултати изискват много големи извадки); конфигурационна грешка замърсява контролната група, а навиците в общата мрежа размиват сравнението — и двете пристрастяват към липса на разлика; една частна градска мрежа с вече високи

стандарти; липса на предварително зададена рамка за не по-ниска ефективност или за безопасност; кратък 14-дневен хоризонт.

Колко уверен трябва да бъде общият читател? Висока увереност, че инструментът не показа сигнал за безопасност в това изпитване и подобри документацията. Висока увереност, че не демонстрира подобрение на 14-дневните пациентски резултати тук. Ниска увереност за твърдение, че „работи“ или „не работи“ като интервенция за полза на пациента — този въпрос действително остава нерешен и изисква много по-голямо изследване. Подходящата позиция е: внимателен реален резултат за инструмент, който не показа сигнал за безопасност и подобри процеса на грижа, не присъда, че AI преобразява — или разрушава — първичната помощ.

Източници

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