The Clean Paper

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

Клінічний AI, який виглядав безпечним і поліпшив документацію — але не результати пацієнтів

Прагматичне кластерно-рандомізоване випробування у 16 кенійських закладах первинної допомоги перевірило генеративний AI-інструмент підтримки рішень, доданий до медичної карти clinical officers. Серед 9 691 пацієнта суворий, експертно підтверджений 14-денний комбінований показник невдачі лікування становив 2,2% з інструментом проти 2,0% без нього (скориговане відношення шансів 0,77; 95% ДІ 0,55–1,08; P = 0,13) — значущої різниці немає. Інструмент не дав сигналу небезпеки, поліпшив документацію за всіма оціненими напрямами, не змінив призначення чи задоволеність і показав тенденцію до нижчих витрат на антибіотики. Це не невдале дослідження, а рідкісне чесне — виміряне на пацієнтах, а не бенчмарках — і воно приходить до неефектної істини: інструмент, який виглядав безпечним і допомагав процесу, не продемонстрував користі для пацієнтів за два тижні, а для виявлення можливої невеликої користі потрібне набагато більше випробування.

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

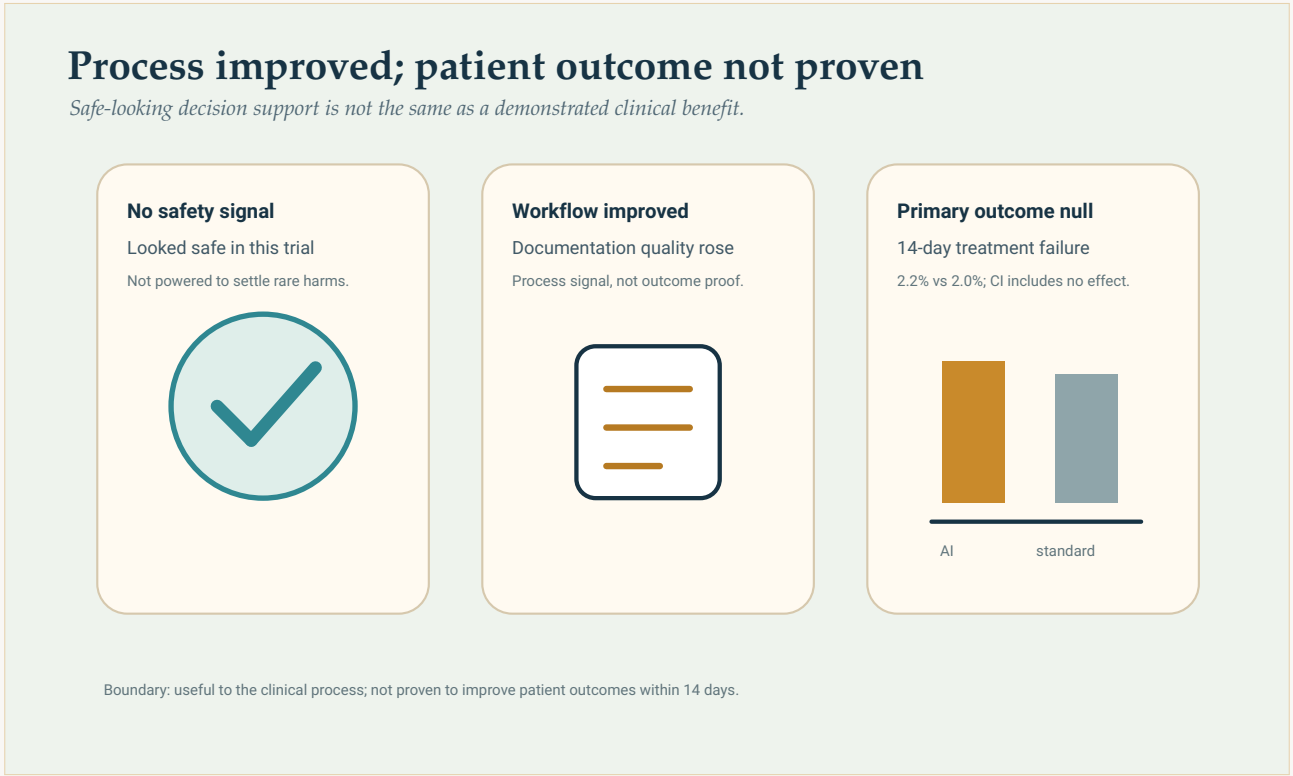
Безпечний і корисний — не те саме, що ефективний

Більшість заголовків про медичний AI виростають із неправильного типу досліджень. Модель добре відповідає на екзаменаційні запитання або перевершує лікарів на спеціально підібраних клінічних віньєтках — і історія пишеться сама: машина готова до клініки.

Це випробування зробило складнішу й рідкіснішу річ. Воно поставило інструмент підтримки рішень на основі генеративного AI у справжню первинну медичну допомогу, у справжні заклади, зі справжніми пацієнтами — і поставило запитання, яке насправді має значення: пацієнтам стало краще?

Чесна відповідь — ні. Принаймні вимірювано й упродовж 14 днів. Інструмент не дав сигналу небезпеки. Він поліпшив якість клінічної документації. Можливо, навіть трохи знизив витрати на деякі ліки. Але значущого зменшення невдач лікування не було, і автори обережно кажуть: якщо користь існує, вона, ймовірно, скромна.

Це не провал дослідження. Це дослідження, яке працює як треба. Так виглядають відповідальні докази щодо AI в медицині, коли вимірюють наслідки для пацієнтів, а не бали бенчмарку.



Інструмент не дав сигналу небезпеки й поліпшив клінічну документацію, але заздалегідь визначений 14-денний результат для пацієнтів значущо не покращився. Допомога процесу — не те саме, що доведена користь для пацієнта.

Що зробили автори

Команда провела прагматичне кластерно-рандомізоване випробування у **16 закладах первинної допомоги** приватної мережі Penda Health у округах Найробі та Кіамбу, Кенія. Допомогу там значною мірою надають **clinical officers** — медичні працівники середнього рівня з трирічним дипломом, часто без легкого доступу до консультації старших лікарів.

Одиницею рандомізації був клініцист, а не пацієнт. **103 clinical officers** випадково розподілили: 52 — у групу втручання, 51 — у контроль. Обидві групи користувалися тією самою хмарною електронною медичною картою. Група втручання додатково отримала **AI Consult** (версія 2.0), інструмент підтримки рішень на основі великої мовної моделі **OpenAI GPT-4o**, інтегрований у цю карту. Він читав інформацію, яку вносив клініцист, і міг вказувати на можливі проблеми з діагнозом чи планом лікування. Остаточне рішення завжди лишалося за клініцистом: пораду можна було прийняти, змінити або проігнорувати.

> **Яка саме модель — і чому деталі важливі**
>
> Інструментом був **AI Consult 2.0** на **OpenAI GPT-4o** (випуск травня 2025 року), доступний через комерційний API OpenAI за корпоративною ліцензією й налаштований на низьку випадковість (temperature 0,1). Він працював усередині спеціальної електронної медичної системи EasyClinic EMR і керувався системними промптами, написаними відповідно до національних клінічних настанов Кенії; автори опублікували повний інструктивний промпт.
> Навіщо така конкретика? Бо результат стосується *однієї конкретної системи* — однієї версії моделі, одного промпту, однієї медичної карти й одного середовища, — а не «LLM у медицині» загалом. Автори підкреслюють те саме: вони називають свій результат *часовим еталоном, а не фіксованою оцінкою можливостей*. Новіша модель, інший промпт або менш цифровізована клініка могли б змінити результат.
> Щодо незалежності: OpenAI пізніше надала підтримку в натуральній формі (кредити на хмарні обчислення й технічні поради щодо API), але автори зазначають, що рішення використовувати OpenAI було ухвалено *до* цієї пропозиції й що OpenAI не брала участі в дизайні випробування, зборі чи аналізі даних або рішенні про публікацію.

Між 22 квітня та 16 липня 2025 року до дослідження включили **9 691 пацієнта**. **Первинний результат навмисно був орієнтований на пацієнта й суворий**: експертно підтверджений *комбінований показник невдачі лікування* протягом 14 днів після візиту. Панель клініцистів, засліплена щодо групи, оцінювала, чи мав пацієнт несприятливий результат — наприклад, хвороба не минула або погіршилася. Випробування зареєстрували заздалегідь (Pan-African Clinical Trials Registry 202502499779176).

Саме цей вибір дизайну є суттю. Легко показати, що AI змінює те, що клініцист записує.

Набагато складніше — і значно важливіше — показати, що він змінює те, що стається з пацієнтом.

Що вони виявили

Первинний результат не покращився. Невдача лікування сталася у 102 із 4 693 пацієнтів (2,2%) у групі AI і у 94 із 4 654 (2,0%) у контролі. Сирі відсотки були трохи вищими з AI, але після коригування точкова оцінка схилялася до користі: **скориговане відношення шансів 0,77 (95% довірчий інтервал 0,55–1,08; P = 0,13)** — статистично незначуще. Те, що напрямок змінюється між сирими й скоригованими числами, не арифметична помилка: коригування враховує відмінності між кластерами клініцистів. У будь-якому разі довірчий інтервал без проблем включає «жодного ефекту», тому користь заявляти не можна, а абсолютний ефект був крихітним.

Як читати відношення шансів, довірчі інтервали й P-значення разом простою мовою, дивіться в посібнику з читання клінічного результату.

Сигналу небезпеки не було — у межах можливостей дослідження. Жодну серйозну небажану подію не визнали пов'язаною з інструментом, а незалежний огляд не знайшов сигналу безпеки. Автори чесно позначають межу такого заспокоєння: випробування не мало достатньої статистичної потужності для рідкісних тяжких шкод і не мало заздалегідь визначеної схеми non-inferiority чи формальної безпекової рамки, тому воно не може *довести* безпеку щодо рідкісних подій.

Документація стала кращою. Серед 2 000 консультацій, переглянутих засліпленими експертами, клініцисти з AI Consult краще документували всі оцінювані компоненти — діагноз, план лікування й загальну повноту запису.

Призначення майже не змінилися. Значущої різниці в призначеннях, зокрема правильному застосуванні антибіотиків, не було (скориговане відношення шансів 0,86; 95% ДІ 0,48–1,55). Частота призначення антибіотиків не змінилася.

Пацієнти різниці не помітили. Серед 826 пацієнтів, які заповнили опитування задоволеності, оцінки були практично однаковими в обох групах; тривалість консультації теж була подібною.

Витрати трохи схилялися вниз. У скоригованому аналізі витрати на антибіотики були нижчими в групі AI — ймовірно, через вибір дешевших препаратів, а не меншу кількість призначень, — і економія на антибіотиках на одного пацієнта виглядала більшою за вартість роботи інструмента на одного пацієнта. Автори позначають це як натяк, а не встановлений факт: повний розрахунок сукупної вартості володіння не входив до випробування.

Однорядкове резюме самих авторів — найточніше: допомога LLM у цих межах виглядала безпечною, але не зменшила невдачі лікування за 14 днів, а будь-яка користь, імовірно, скромна.

Чому «немає значущої різниці» не означає «це не працює»

Нульовий первинний результат легко перебільшити в будь-який бік. Дві речі не дозволяють простій історії.

По-перше, **випробування було спроєктоване, щоб виявити більший ефект, ніж той, який воно побачило**. Серйозні несприятливі результати в первинній допомозі рідкісні — тут близько 2%, — тому для відділення невеликої реальної користі від шуму потрібні величезні вибірки. Постфактум розрахунки потужності авторів свідчать, що для виявлення ефекту такого розміру, як спостережений, знадобилося б **набагато більше випробування — порядку понад 100 000 пацієнтів**. Незначущий результат у дослідженні цього масштабу не виключає невелику реальну користь; він означає, що це дослідження не могло її розрізнити.

По-друге, **порівняння частково розмилося**. Через помилку конфігурації деякі клініцисти контрольної групи короткий час мали доступ до AI Consult, а працівники однієї мережі спілкуються й переносять звички через межу між групами. Обидва ефекти роблять групи подібнішими й тягнуть будь-яку справжню різницю до нуля. Крім того, сама мережа вже працювала за відносно високими стандартами, тож інструмент мав менше простору для поліпшення.

Ніщо з цього не рятує заголовок про «прорив». Але правильне читання має бути каліброваним, а не зневажливим: за найскладнішим і найчеснішим показником цей інструмент не продемонстрував користі для пацієнтів за два тижні — водночас він помітно допоміг із документацією й виглядав безпечним.

Чого це *не* доводить

- Це **не показує**, що AI покращив результати пацієнтів. За первинним 14-денним показником значущої користі не було.
- Це **не показує**, що AI марний. Точкова оцінка була на його користь, документація покращилася, витрати на ліки схилялися вниз; нульовий результат сумісний із невеликою реальною користю, яку дослідження було замалим для підтвердження.
- Це **не доводить безпеку щодо рідкісних шкод**. Сигналу небезпеки не знайшли, але дослідження не було спроєктоване чи достатньо потужне, щоб сертифікувати безпеку для нечастих тяжких подій.
- Це **не показує, що «AI перемагає лікарів» або замінює клініцистів**. Це *підтримка* рішення; остаточну владу прийняти чи відхилити пораду мав клініцист.
- Результат **не узагальнюється автоматично**. Випробування відбулося в одній приватній міській мережі Кенії; у сільських, приміських чи багатших системах результати можуть відрізнятися в будь-який бік.
- Воно **не встановлює економію коштів**. Сигнал витрат цікавий, але це не завершена економічна оцінка.

Наскільки сильні докази?

Для центрального твердження — що доведеного зменшення короткострокової шкоди пацієнтам немає — дизайн доказів сильний, а висновок належно скромний. Проспективне, заздалегідь зареєстроване кластерно-рандомізоване випробування із засліпленим комбінованим показником на рівні пацієнта — майже найкращий тип реальних доказів, який можна отримати для такого інструмента. Воно набагато інформативніше за бал бенчмарку чи дослідження клінічних віньєток.

Для вторинних результатів — кращої документації, незмінних призначень, схожої задоволеності, нижчих витрат на антибіотики — докази добрі, але їх слід читати як вторинні: підтримувальні сигнали, не головний висновок, і вони підлягають тим самим обмеженням через контамінацію груп та одну мережу.

Для безпеки докази заспокоюють, але мають чітку межу: сигналу немає, однак це не було дослідження рідкісної шкоди.

Найкорисніша позиція — не «працює» й не «провалилося». А така: ретельне реальне випробування показало, що цей AI-інструмент не дав сигналу небезпеки й поліпшив процес надання допомоги, але не продемонстрував користі для пацієнтів упродовж двох тижнів — і для виявлення такої користі, якщо вона є, потрібне значно більше дослідження.

Чому це важливо

Дебату про медичний AI відчайдушно бракує саме таких доказів. Є тисячі статей, де моделі блискуче складають іспити або дорівнюють клініцистам на акуратних задачах. Великих прагматичних рандомізованих випробувань, що вимірюють, чи стало реальним пацієнтам краще, — дуже мало. Це одне з них, і воно приходить до неефектної істини: скласти тест — не те саме, що допомогти пацієнту.

Цей розрив і є всією історією. Інструмент може бути справді корисним клініцистам — кращі записи, дешевші ліки, ще одна пара очей — і водночас не зрушити жорсткий клінічний результат за два тижні. Обидва факти можуть бути істинними одночасно, і зріла система охорони здоров’я має втримати їх разом, а не вибрати зручний.

Робота також корисно пересуває тягар доказу. Якщо компанія хоче заявити, що її клінічний AI покращує допомогу, релевантний доказ — не таблиця лідерів. Це випробування на кшталт цього, з результатами, важливими для пацієнтів, — і в ідеалі ще більше, бо чесний урок тут у тому, що скромну користь видно лише у великих дослідженнях.

Короткий підсумок

Прагматичне кластерно-рандомізоване випробування у 16 кенійських закладах первинної допомоги перевірило генеративний AI-інструмент підтримки рішень AI Consult, доданий до електронної карти clinical officers. Серед 9 691 пацієнта експертно підтверджений комбінований показник невдачі лікування протягом 14 днів становив 2,2% з інструментом проти 2,0% без нього (скориговане відношення шансів 0,77; 95% ДІ 0,55–1,08; P = 0,13) — значущої різниці немає. Інструмент не дав сигналу небезпеки, поліпшив клінічну документацію за всіма оціненими напрямами, не змінив призначення, не вплинув на задоволеність пацієнтів і асоціювався з дещо нижчими витратами на антибіотики. Автори роблять висновок: у цих межах він виглядав безпечним, але не зменшив невдачі лікування; можлива користь, імовірно, скромна, і для ефекту спостереженого розміру знадобилося б набагато більше випробування — порядку 100 000 пацієнтів. Результат не показує ані того, що клінічний AI покращує результати пацієнтів, ані того, що він марний: він показує, що інструмент, який виглядав безпечним і допомагав процесу, не продемонстрував користі для пацієнтів за два тижні в одній приватній міській мережі.

Тверезa оцінка

Що показує стаття: У реальному рандомізованому випробуванні додавання LLM-інструмента підтримки рішень до первинної допомоги не дало сигналу небезпеки, поліпшило якість клінічної документації, не змінило призначення й не зменшило значущо суворий 14-денний комбінований показник невдачі лікування.

Що правдоподібне, але не доведене: Що інструмент дає невелике реальне зниження невдач лікування, надто мале для цього випробування; що він економить гроші після врахування всіх витрат; що краща документація зрештою перетворюється на кращу допомогу.

Чого робота не показує: Що клінічний AI покращує результати пацієнтів; що він безпечний щодо рідкісних подій; що він замінює чи перевершує клініцистів; що результати переносяться на сільські або багатші системи; що економія коштів встановлена.

Основні обмеження: Потужність розраховувалася на більший ефект, ніж спостережений; рідкісні результати потребують дуже великих вибірок. Помилка конфігурації контамінувала контрольну групу, а спільна мережа розмиває відмінності — обидва фактори зміщують результат до нуля. Одна приватна міська мережа з уже високими стандартами; відсутність заздалегідь визначеної non-inferiority або безпекової рамки; короткий 14-денний горизонт.

Скільки довіри варто мати звичайному читачеві? Високу, що в цьому випробуванні інструмент не дав сигналу небезпеки й покращив документацію. Високу, що тут

він не продемонстрував покращення 14-денних результатів пацієнтів. Низьку щодо будь-якого загального вердикту «працює» чи «не працює» як втручання для користі пацієнтів — це питання справді не закрите й потребує значно більшого дослідження. Доречна позиція: ретельний реальний результат про інструмент, що не дав сигналу небезпеки й поліпшив процес допомоги, а не вирок, що AI перетворює — або руйнує — первинну медицину.

Джерела

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