The Clean Paper

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

Клінічны AI, які выглядаў бяспечным і палепшыў дакументацыю — але не вынікі пацыентаў

Прагматычнае кластарна-рандомізаванае выпрабаванне ў 16 кенійскіх установах першаснай дапамогі праверыла генератыўны AI-інструмент падтрымкі рашэнняў, дададзены ў медыцынскую карту clinical officers. Сярод 9 691 пацыента строгі, пацверджаны экспертамі 14-дзённы камбінаваны паказчык няўдачы лячэння склаў 2,2% з інструментам супраць 2,0% без яго (скарэкціраваная адносіна шанцаў 0,77; 95% ДІ 0,55–1,08; P = 0,13) — значнай розніцы няма. Інструмент не даў сігналу небяспекі, палепшыў дакументацыю па ўсіх ацэненых кірунках, не змяніў прызначэнні або задаволенасць і паказаў тэндэнцыю да меншых выдаткаў на антыбіётыкі. Гэта не няўдалае даследаванне, а рэдкае сумленнае — вымеранае на пацыентах, а не бенчмарках — і яно прыходзіць да неэфектнай ісціны: інструмент, які выглядаў бяспечным і дапамагаў працэсу, не прадэманстраваў карысці для пацыентаў за два тыдні, а для выяўлення магчымай невялікай карысці спатрэбіцца значна большае выпрабаванне.

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

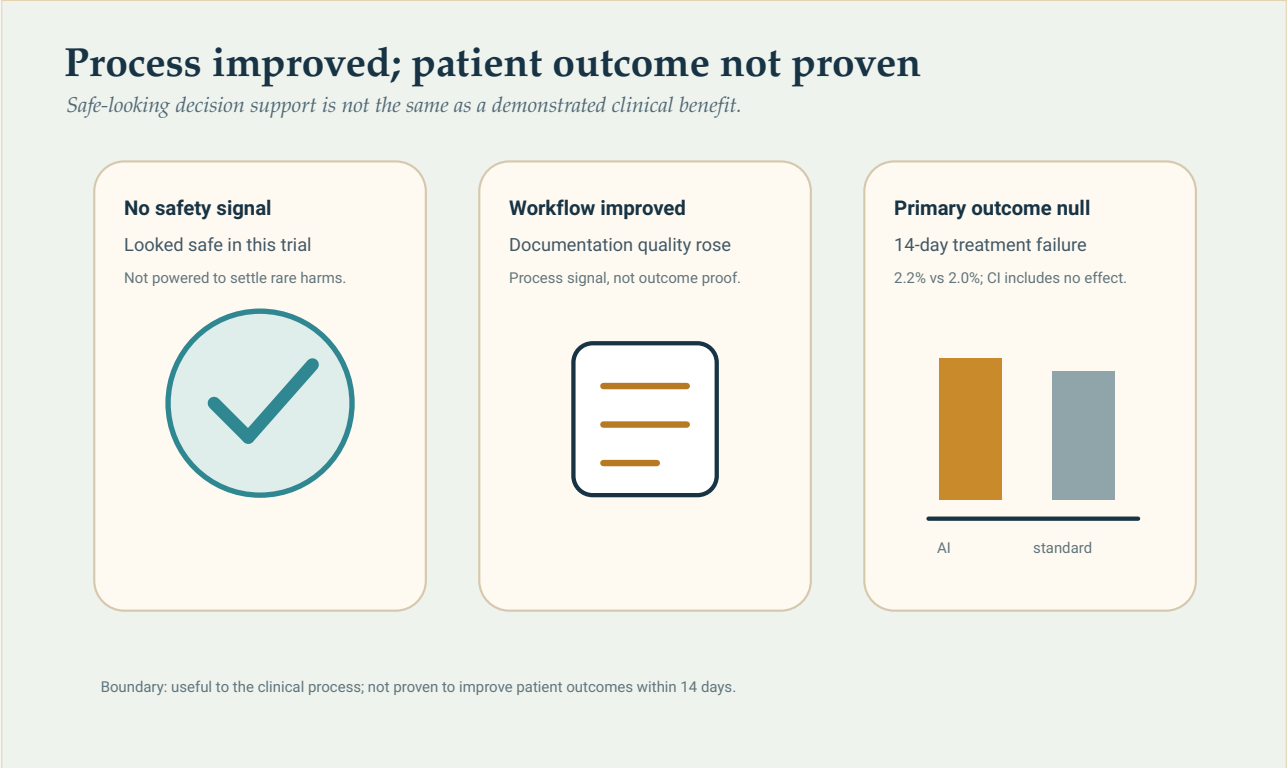
Бяспечны і карысны — не тое самае, што эфектыўны

Большасць загалоўкаў пра медыцынскі AI вырастаюць з няправільнага тыпу даследаванняў. Мадэль добра адказвае на экзаменацыйныя пытанні або пераўзыходзіць лекараў на спецыяльна падабраных клінічных віньетках — і гісторыя пішацца сама: машына гатовая да клінікі.

Гэтае выпрабаванне зрабіла больш цяжкую і рэдкую рэч. Яно змясціла інструмент падтрымкі рашэнняў на аснове генератыўнага AI у сапраўдную першасную медыцынскую дапамогу, у сапраўдныя ўстановы, з сапраўднымі пацыентамі — і паставіла пытанне, якое сапраўды мае значэнне: пацыентам стала лепш?

Сумленны адказ — не. Прынамсі вымерна і на працягу 14 дзён. Інструмент не даў сігналу небяспекі. Ён палепшыў якасць клінічнай дакументацыі. Магчыма, нават крыху знізіў выдаткі на некаторыя лекі. Але значнага памяншэння няўдач лячэння не было, і аўтары асцярожна кажуць: калі карысць існуе, яна, верагодна, сціплая.

Гэта не правал даследавання. Гэта даследаванне, якое працуе як трэба. Так выглядаюць адказныя доказы пра AI у медыцыне, калі вымяраюць наступствы для пацыентаў, а не балы бенчмарка.



Інструмент не даў сігналу небяспекі і палепшыў клінічную дакументацыю, але загадзя вызначаны 14-дзённы вынік для пацыентаў значна не палепшыўся. Дапамога працэсу — не тое самае, што даказаная карысць для пацыента.

Што зрабілі аўтары

Каманда правяла прагматычнае кластарна-рандомізаванае выпрабаванне ў **16 установах першаснай дапамогі** прыватнай сеткі Penda Health у акругах Найробі і Кіямбу, Кенія. Дапамогу там у значнай ступені аказваюць **clinical officers** — медыцынскія работнікі сярэдняга ўзроўню з трохгадовым дыпломам, часта без лёгкага доступу да кансультацыі старэйшых лекараў.

Адзінкай рандомізацыі быў клініцыст, а не пацыент. **103 clinical officers** выпадкова размеркавалі: 52 — у групу ўмяшання, 51 — у кантроль. Абедзве групы карысталіся той самай воблачнай электроннай медыцынскай картай. Група ўмяшання дадаткова атрымала **AI Consult** (версія 2.0), інструмент падтрымкі рашэнняў на аснове вялікай моўнай мадэлі **OpenAI GPT-4o**, інтэграваны ў гэтую карту. Ён чытаў інфармацыю, якую ўносіў клініцыст, і мог паказваць на магчымыя праблемы з дыягназам або планам лячэння. Канчатковае рашэнне заўсёды заставалася за клініцыстам: параду можна было прыняць, змяніць або праігнараваць.

> **Якая менавіта мадэль — і чаму дэталі важныя**
>
> Інструментам быў **AI Consult 2.0** на **OpenAI GPT-4o** (выпуск мая 2025 года), даступны праз камерцыйны API OpenAI паводле карпаратыўнай ліцэнзіі і настроены на нізкую выпадковасць (temperature 0,1). Ён працаваў у спецыяльнай электроннай медыцынскай сістэме EasyClinic EMR і кіраваўся сістэмнымі промптамі, напісанымі ў адпаведнасці з нацыянальнымі клінічнымі рэкамендацыямі Кеніі; аўтары апублікавалі поўны інструктыўны промпт.
>
> Навошта такая канкрэтнасць? Бо вынік адносіцца да *адной канкрэтнай сістэмы* — адной версіі мадэлі, аднаго промпта, адной медыцынскай карты і аднаго асяроддзя, — а не да «LLM у медыцыне» ўвогуле. Аўтары падкрэсліваюць тое самае: яны называюць свой вынік *часовым эталонам, а не фіксаванай ацэнкай магчымасцей*. Навейшая мадэль, іншы промпт або менш лічбавізаваная клініка маглі б змяніць вынік.
>
> Што да незалежнасці: OpenAI пазней дала падтрымку ў натуральнай форме (крэдыты на воблачныя вылічэнні і тэхнічныя парады па API), але аўтары адзначаюць, што рашэнне выкарыстоўваць OpenAI было прынята *да* гэтай прапановы і што OpenAI не ўдзельнічала ў дызайне выпрабавання, зборы або аналізе даных ці рашэнні пра публікацыю.

Паміж 22 красавіка і 16 ліпеня 2025 года ў даследаванне ўключылі **9 691 пацыента**. **Першасны вынік наўмысна быў арыентаваны на пацыента і строгі**: пацверджаны экспертамі *камбінаваны паказчык няўдачы лячэння* на працягу 14 дзён пасля візіту. Панэль клініцыстаў, аслепленая адносна групы, ацэньвала, ці быў у пацыента неспрыяльны вынік — напрыклад, хвароба не прайшла або пагоршылася. Выпрабаванне загадзя зарэгістравалі (Pan-African Clinical Trials Registry 202502499779176).

Менавіта гэты выбар дызайну — сутнасць. Лёгка паказаць, што AI змяняе тое, што

клініцыст запісвае. Значна цяжэй — і нашмат важней — паказаць, што ён змяняе тое, што адбываецца з пацыентам.

Што яны выявілі

Першасны вынік не палепшыўся. Няўдача лячэння адбылася ў 102 з 4 693 пацыентаў (2,2%) у групе AI і ў 94 з 4 654 (2,0%) у кантролі. Сырыя працэнты былі крыху вышэйшыя з AI, але пасля карэкціроўкі кропкавая ацэнка схілялася да карысці: **скарэкціраваная адносіна шанцаў 0,77 (95% даверны інтэрвал 0,55–1,08; P = 0,13)** — статыстычна нязначна. Тое, што кірунак мяняецца паміж сырымі і скарэкціраванымі лікамі, не арыфметычная памылка: карэкціроўка ўлічвае адрозненні паміж кластарамі клініцыстаў. У любым выпадку даверны інтэрвал без праблем уключае «ніякага эфекту», таму заяўляць пра карысць нельга, а абсалютны эфект быў мізэрным.

Як чытаць адносіну шанцаў, даверныя інтэрвалы і P-значэнні разам простай мовай, глядзіце ў кіраўніцтве па чытанні клінічнага выніку.

Сігналу небяспекі не было — у межах магчымасцей даследавання. Ніводную сур'ёзную непажаданую з'яву не прызналі звязанай з інструментам, а незалежны агляд не знайшоў сігналу бяспекі. Аўтары сумленна пазначаюць мяжу такога супакаення: выпрабаванне не мела дастатковай статыстычнай магутнасці для рэдкіх цяжкіх шкод і не мела загадзя вызначанай схемы non-inferiority або фармальнай рамкі бяспекі, таму яно не можа *даказаць* бяспеку адносна рэдкіх падзей.

Дакументацыя стала лепшай. Сярод 2 000 кансультацый, прагледжаных аслепленымі экспертамі, клініцысты з AI Consult лепш дакументавалі ўсе ацэненыя кампаненты — дыягназ, план лячэння і агульную паўнату запісу.

Прызначэнні амаль не змяніліся. Значнай розніцы ў прызначэннях, уключаючы правільнае ўжыванне антыбіётыкаў, не было (скарэкціраваная адносіна шанцаў 0,86; 95% ДІ 0,48–1,55). Частата прызначэння антыбіётыкаў не змянілася.

Пацыенты розніцы не заўважылі. Сярод 826 пацыентаў, якія запоўнілі апытанне задаволенасці, ацэнкі былі практычна аднолькавымі ў абедзвюх групах; працягласць кансультацыі таксама была падобнай.

Выдаткі крыху схіляліся ўніз. У скарэкціраваным аналізе выдаткі на антыбіётыкі былі ніжэйшыя ў групе AI — верагодна, праз выбар таннейшых прэпаратаў, а не меншую колькасць прызначэнняў, — і эканомія на антыбіётыках на аднаго пацыента выглядала большай за кошт працы інструмента на аднаго пацыента. Аўтары пазначаюць гэта як намёк, а не ўсталяваны факт: поўны разлік сукупнага кошту валодання не ўваходзіў у выпрабаванне.

Аднарадковае рэзюмэ саміх аўтараў — самае чыстае: дапамога LLM у гэтых межах выглядала бяспечнай, але не паменшыла няўдачы лячэння за 14 дзён, а любая карысць, верагодна, сціплая.

Чаму «няма значнай розніцы» не азначае «гэта не працуе»

Нулявы першасны вынік лёгка перабольшыць у любы бок. Дзве рэчы не дазваляюць простай гісторыі.

Па-першае, **выпрабаванне было спраектавана для выяўлення большага эфекту, чым той, які яно ўбачыла**. Сур'ёзныя неспрыяльныя вынікі ў першаснай дапамозе рэдкія — тут каля 2%, — таму для аддзялення невялікай рэальнай карысці ад шуму патрэбныя велізарныя выбаркі. Постфактум разлікі магутнасці аўтараў паказваюць, што для выяўлення эфекту такога памеру, як назіраны, спатрэбілася б **значна большае выпрабаванне — парадку больш за 100 000 пацыентаў**. Нязначны вынік у даследаванні такога маштабу не выключае невялікай рэальнай карысці; ён азначае, што гэтае даследаванне не магло яе адрозніць.

Па-другое, **параўнанне часткова размылася**. Праз памылку канфігурацыі некаторыя клініцысты кантрольнай групы кароткі час мелі доступ да AI Consult, а супрацоўнікі адной сеткі размаўляюць і пераносяць звычкі праз мяжу паміж групамі. Абодва эфекты робяць групы больш падобнымі і цягнуць любую сапраўдную розніцу да нуля. Акрамя таго, сама сетка ўжо працавала паводле адносна высокіх стандартаў, таму ў інструмента было менш прасторы для паляпшэння.

Нішто з гэтага не ратуе загаловак пра «прарыў». Але правільнае прачытанне павінна быць адкалібраваным, а не грэблівым: паводле самага складанага і сумленнага паказчыка гэты інструмент не прадэманстраваў карысці для пацыентаў за два тыдні — і адначасова прыкметна дапамог з дакументацыяй і выглядаў бяспечным.

Чаго гэта *не* даказвае

- Гэта **не паказвае**, што AI палепшыў вынікі пацыентаў. Паводле першаснага 14-дзённага паказчыка значнай карысці не было.
- Гэта **не паказвае**, што AI бескарысны. Кропкавая ацэнка была на яго карысць, дакументацыя палепшылася, выдаткі на лекі схіляліся ўніз; нулявы вынік сумяшчальны з невялікай рэальнай карысцю, для пацвярджэння якой даследаванне было занадта малым.
- Гэта **не даказвае бяспеку адносна рэдкіх шкод**. Сігналу небяспекі не знайшлі, але даследаванне не было спраектавана або дастаткова магутнае, каб сертыфікаваць бяспеку для нячастых цяжкіх падзей.
- Гэта **не паказвае, што «AI перамагае лекараў» або замяняе клініцыстаў**. Гэта *падтрымка* рашэння; канчатковая ўлада прыняць або адхіліць параду была ў клініцыста.
- Вынік **не абагульняецца аўтаматычна**. Выпрабаванне праходзіла ў адной прыватнай гарадской сетцы Кеніі; у сельскіх, прыгарадных або больш заможных сістэмах вынікі могуць адрозніваца ў любы бок.

- Яно **не ўсталёўвае эканомію сродкаў**. Сігнал выдаткаў цікавы, але гэта не завершаная эканамічная ацэнка.

Наколькі моцныя доказы?

Для цэнтральнага сцвярджэння — што даказанага памяншэння кароткатэрміновай шкоды пацыентам няма — дызайн доказаў моцны, а выснова належным чынам сціплая. Праспектыўнае, загадзя зарэгістраванае кластарна-рандомізаванае выпрабаванне з аслепленым камбінаваным паказчыкам на ўзроўні пацыента — амаль найлепшы тып рэальных доказаў, які можна атрымаць для такога інструмента. Яно нашмат інфарматыўнейшае за бал бенчмарка або даследаванне клінічных віньетак.

Для другасных вынікаў — лепшай дакументацыі, нязменных прызначэнняў, падобнай задаволенасці, меншых выдаткаў на антыбіётыкі — доказы добрыя, але іх варта чытаць як другасныя: падтрымліваючыя сігналы, а не галоўную выснову, і яны падпадаюць пад тыя самыя абмежаванні праз кантамінацыю груп і адну сетку.

Для бяспекі доказы супакойваюць, але маюць выразную мяжу: сігналу няма, аднак гэта не было даследаваннем рэдкай шкоды.

Самая карысная пазіцыя — не «працуе» і не «правалілася». Яна такая: стараннае рэальнае выпрабаванне паказала, што гэты AI-інструмент не даў сігналу небяспекі і палепшыў працэс аказання дапамогі, але не прадэманстраваў карысці для пацыентаў на працягу двух тыдняў — і для выяўлення такой карысці, калі яна ёсць, спатрэбіцца значна большае даследаванне.

Чаму гэта важна

Дэбатам пра медыцынскі AI адчайна не хапае менавіта такіх доказаў. Ёсць тысячы артыкулаў, дзе мадэлі бліскуча здаюць экзамены або раўняюцца з клініцыстамі на акуратных задачах. Вялікіх прагматычных рандомізаваных выпрабаванняў, якія вымяраюць, ці стала рэальным пацыентам лепш, — вельмі мала. Гэта адно з іх, і яно прыходзіць да неэфектнай ісціны: здаць тэст — не тое самае, што дапамагчы пацыенту.

Гэты разрыў і ёсць уся гісторыя. Інструмент можа быць сапраўды карысным клініцыстам — лепшыя запісы, таннейшыя лекі, яшчэ адна пара вачэй — і адначасова не зрушыць жорсткі клінічны вынік за два тыдні. Абодва факты могуць быць праўдзівымі адначасова, і спелая сістэма аховы здароўя павінна трымаць іх разам, а не выбіраць зручны.

Праца таксама карысна зрушвае цяжар доказу. Калі кампанія хоча заявіць, што яе клінічны AI паляпшае дапамогу, рэлевантны доказ — не табліца лідараў. Гэта выпрабаванне накшталт гэтага, з вынікамі, важнымі для пацыентаў, — і ў ідэале яшчэ большае, бо сумленны ўрок тут у тым, што сціплую карысць відаць толькі ў вялікіх

даследаваннях.

Коратка

Прагматычнае кластарна-рандомізаванае выпрабаванне ў 16 кенійскіх установах першаснай дапамогі праверыла генератыўны AI-інструмент падтрымкі рашэнняў AI Consult, дададзены ў электронную карту clinical officers. Сярод 9 691 пацыента пацверджаны экспертамі камбінаваны паказчык няўдачы лячэння на працягу 14 дзён склаў 2,2% з інструментам супраць 2,0% без яго (скарэкціраваная адносіна шанцаў 0,77; 95% ДІ 0,55–1,08; P = 0,13) — значнай розніцы няма. Інструмент не даў сігналу небяспекі, палепшыў клінічную дакументацыю па ўсіх ацэненых кірунках, не змяніў прызначэнні, не паўплываў на задаволенасць пацыентаў і асацыяваўся з крыху меншымі выдаткамі на антыбіётыкі. Аўтары робяць выснову: у гэтых межах ён выглядаў бяспечным, але не паменшыў няўдачы лячэння; магчымая карысць, верагодна, сціплая, і для эфекту назіранага памеру спатрэбілася б значна большае выпрабаванне — парадку 100 000 пацыентаў. Вынік не паказвае ні таго, што клінічны AI паляпшае вынікі пацыентаў, ні таго, што ён бескарысны: ён паказвае, што інструмент, які выглядаў бяспечным і дапамагаў працэсу, не прадэманстраваў карысці для пацыентаў за два тыдні ў адной прыватнай гарадской сетцы.

Праверка без ажыятажу

Што паказвае артыкул: У рэальным рандомізаваным выпрабаванні даданне LLM-інструмента падтрымкі рашэнняў да першаснай дапамогі не дало сігналу небяспекі, палепшыла якасць клінічнай дакументацыі, не змяніла прызначэнні і не паменшыла значна строгі 14-дзённы камбінаваны паказчык няўдачы лячэння.

Што праўдападобна, але не даказана: Што інструмент дае невялікае рэальнае зніжэнне няўдач лячэння, занадта малое для гэтага выпрабавання; што ён эканоміць грошы пасля ўліку ўсіх выдаткаў; што лепшая дакументацыя ўрэшце ператвараецца ў лепшую дапамогу.

Чаго праца не паказвае: Што клінічны AI паляпшае вынікі пацыентаў; што ён бяспечны адносна рэдкіх падзей; што ён замяняе або пераўзыходзіць клініцыстаў; што вынікі пераносяцца на сельскія або больш заможныя сістэмы; што эканомія сродкаў установлена.

Асноўныя абмежаванні: Магутнасць разлічвалася на большы эфект, чым назіраны; рэдкія вынікі патрабуюць вельмі вялікіх выбарак. Памылка канфігурацыі кантамінавала кантрольную групу, а агульная сетка размывае адрозненні — абодва фактары зрушваюць вынік да нуля. Адна прыватная гарадская сетка з ужо высокімі стандартамі; адсутнасць загадзя вызначанай non-inferiority або рамкі бяспекі; кароткі 14-дзённы гарызонт.

Колькі даверу варта мець звычайнаму чытачу? Высокага — да таго, што ў гэтым выпрабаванні інструмент не даў сігналу небяспекі і палепшыў дакументацыю. Высокага — да таго, што тут ён не прадэманстраваў паляпшэння 14-дзённых вынікаў пацыентаў. Нізкага — адносна любога агульнага вердыкту «працуе» або «не працуе» як умяшанне для карысці пацыентаў: гэтае пытанне сапраўды не закрыта і патрабуе значна большага даследавання. Дарэчная пазіцыя: старанны рэальны вынік пра інструмент, які не даў сігналу небяспекі і палепшыў працэс дапамогі, а не прысуд, што AI пераўтворыць — або разбурыць — першасную медыцыну.

Крыніцы

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