



The CLEAN PAPER

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

Klinikinis DI atrodė saugus ir pagerino dokumentaciją, bet pacientų baigčių nepagerino

Pragmatiniame klasteriniame randomizuotame tyrime 16 Kenijos pirminės priežiūros įstaigų išbandyta generatyvinio DI sprendimų palaikymo priemonė. Tarp 9 691 paciento griežtas, ekspertų vertintas sudėtinis gydymo nesėkmės per 14 dienų rodiklis buvo 2,2 % su priemone ir 2,0 % be jos (koreguotas šansų santykis 0,77, 95 % PI 0,55–1,08, $P = 0,13$) – reikšmingo skirtumo nebuvo. Priemonė nekėlė saugos signalo, pagerino dokumentaciją visose vertintose srityse, nepakeitė vaistų skyrimo ar pasitenkinimo ir turėjo mažesnių antibiotikų išlaidų tendenciją. Tai ne nesėkmingas tyrimas, o retas ir sąžiningas – matuojantis pacientus, ne etalonus. Jis parodė nepatrauklią, bet svarbią tiesą: priemonė atrodė saugi ir padėjo procesui, tačiau per dvi savaites išmatuojamos naudos pacientams neparodė; tokiai nedidelei naudai aptikti reikėtų gerokai didesnio tyrimo.

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

Saugus ir naudingas procesui dar nereikia veiksmingo pacientui

Daugelis medicininio DI antraščių remiasi netinkamo tipo tyrimais. Modelis gerai atsako į egzamino klausimus ar pranoksta gydytojus kruopščiai atrinktose klinikinėse situacijose, ir istorija parašoma pati: mašina pasirengusi klinikai.

Šis tyrimas pasirinko sunkesnę ir retesnę kelią. Generatyvinio DI sprendimų palaikymo priemonė buvo įdiegta realioje pirminėje sveikatos priežiūroje, realiose įstaigose ir su realiais pacientais, o tada užduotas klausimas, kuris iš tiesų svarbus: ar pacientams sekėsi geriau?

Sąžiningas atsakymas – ne, bent jau ne išmatuojamai per 14 dienų. Tyrime nepasirodė saugos signalas. Pagerėjo klinikinė dokumentacija. Gali būti, kad sumažėjo kai kurios vaistų išlaidos. Tačiau gydymo nesėkmių skaičius reikšmingai nesumažėjo, o autoriai atsargiai pažymi, kad jei nauda egzistuoja, ji tikriausiai nedidelė.

Tai nėra tyrimo nesėkmė. Tai tyrimas, atliekantis savo darbą. Taip atrodo atsakingi medicininio DI įrodymai, kai vertinami pacientai, o ne etalonų balai.

Process improved; patient outcome not proven

Safe-looking decision support is not the same as a demonstrated clinical benefit.

No safety signal

Looked safe in this trial

Not powered to settle rare harms.



Workflow improved

Documentation quality rose

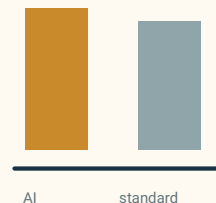
Process signal, not outcome proof.



Primary outcome null

14-day treatment failure

2.2% vs 2.0%; CI includes no effect.



Boundary: useful to the clinical process; not proven to improve patient outcomes within 14 days.

Tyrime nebuvo aptikta saugos signalo ir pagerėjo klinikinė dokumentacija, tačiau iš anksto numatytas 14 dienų paciento baigties rodiklis reikšmingai nepagerėjo. Geresnis procesas nėra tas pats, kas įrodytas pacientų rezultato pagerėjimas.

Ką padarė autoriai

Komanda atliko pragmatinį klasterinį randomizuotą tyrimą **16 pirminės sveikatos priežiūros įstaigų**, priklausančių privačiam Penda Health tinklui Nairobi ir Kiambu apskrityse Kenijoje. Daugumą priežiūros ten teikia **clinical officers** – vidurinės grandies klinicianai, baigę trejų metų profesinę programą ir dažnai neturintys lengvos prieigos prie vyresnio specialisto konsultacijos.

Randomizavimo vienetas buvo klinicianas, ne pacientas. Atsitiktinai paskirstyti **103 clinical officers**: 52 į intervencijos grupę ir 51 į kontrolę. Abi grupės naudojo tą pačią debesijos elektroninę medicinos įrašų sistemą. Intervencijos grupę papildomai turėjo **AI Consult** (2.0 versiją) – sprendimų palaikymo priemonę, sukurtą naudojant **OpenAI GPT-4o** didįjį kalbos modelį ir integruotą į patį įrašą. Sistema skaitė tai, ką klinicianas dokumentavo, ir galėjo pažymėti galimas diagnozės ar gydymo plano problemas. Klinicianai išlaikė visą sprendimo teisę: pasiūlymus galėjo priimti, pakeisti arba ignoruoti.

Koks modelis buvo naudojamas ir kodėl detalės svarbios?

Priemonė buvo **AI Consult 2.0**, veikianti su **OpenAI GPT-4o** (2025 m. gegužės leidimu), naudojamu per komercinę OpenAI API pagal įmonės licenciją ir nustatytą mažą atsitiktinumą (temperature 0.1). Ji buvo integruota į specialią elektroninių įrašų sistemą EasyClinic EMR ir valdoma sisteminėmis instrukcijomis, suderintomis su Kenijos nacionalinėmis gydymo gairėmis; autoriai paskelbė visą instrukcijų tekstą.

Kodėl tai reikia nurodyti? Nes rezultatas galioja *vienai konkrečiai sistemai* – vienai modelio versijai, vienam instrukcijų rinkiniui, vienai įrašų sistemai ir vienai aplinkai – o ne „LLM medicinoje“ apskritai. Autoriai pabrėžia tą patį: savo rezultatą jie vadina *laiko momentą fiksuojančiu etalonu, o ne nekintamu gebėjimų įverčiu*. Naujesnis modelis, kitokia instrukcija ar mažiau skaitmenizuota klinika gali duoti kitokį rezultatą.

Dėl nepriklausomumo: vėliau OpenAI suteikė paramą natūra – debesijos skaičiavimo kreditų ir techninių patarimų dėl API – tačiau autoriai teigia, kad sprendimas naudoti OpenAI buvo priimtas *dar prieš* šį pasiūlymą, o bendrovė nedalyvavo kuriant tyrimą, renkant ar analizuojant duomenis ir sprendžiant dėl publikavimo.

Nuo 2025 m. balandžio 22 d. iki liepos 16 d. įtrauktas **9 691 pacientas**. **Pirminė baigtis buvo sąmoningai orientuota į pacientą ir griežta**: ekspertų, nežinojusių tyrimo grupės, įvertintas *sudėtinis gydymo nesėkmės rodiklis* per 14 dienų po vizito – klinicianų komisija sprendė, ar pacientui pasireiškė nepalanki baigtis, pavyzdžiui, liga neišnyko arba pablogėjo. Tyrimas iš anksto registruotas Pan-African Clinical Trials Registry (202502499779176).

Būtent šis dizaino pasirinkimas yra esmė. Lengva parodyti, kad DI keičia tai, ką klinicianas užrašo. Daug sunkiau – ir daug prasmingiau – parodyti, kad pasikeičia tai, kas nutinka pacientui.

Ką jie nustatė

Pirminė baigtis nepagerėjo. Gydomo nesėkmė pasireiškė 102 iš 4 693 pacientų (2,2 %) DI grupėje ir 94 iš 4 654 (2,0 %) kontrolėje. Neapdoroti procentai buvo truputį didesni DI grupėje, bet po koregavimo taškinis įvertis krypo naudos link: **koreguotas šansų santykis 0,77 (95 % pasikliautinis intervalas 0,55–1,08, P = 0,13)** – statistiškai nereikšminga. Skirtumas tarp neapdorotų ir koreguotų skaičių nėra aritmetinė klaida; korekcija atsižvelgia į skirtumus tarp klinikistų klasterių. Bet kuriuo atveju pasikliautinis intervalas aiškiai apima „jokio poveikio“ reikšmę, todėl naudos teigti negalima, o absoliutus skirtumas labai mažas.

Kaip kartu skaityti šansų santykį, pasikliautinąjį intervalą ir P reikšmę, žr. klinikinių rezultatų skaitymo gidą.

Saugos signalo nebuvo – bet tik šio tyrimo ribose. Nė vienas sunkus nepageidaujamas įvykis nebuvo priskirtas priemonei, o nepriklausoma peržiūra saugos signalo nerado. Autoriai aiškiai nurodo ribą: tyrimas neturėjo pakankamos statistinės galios aptikti retus sunkius pakenkimus ir neturėjo iš anksto numatyto ne mažesnio veiksmingumo (*noninferiority*) ar formalaus saugos dizaino, todėl negali įrodyti saugumo dėl retų reiškinių.

Dokumentacija pagerėjo. Iš 2 000 aklaai ekspertų įvertintų vizitų DI naudoję klinikistai geriau dokumentavo visas vertintas sritis – diagnozę, gydymo planą ir bendrą įrašo išsamumą.

Vaistų skyrimas beveik nepasikeitė. Reikšmingo skirtumo skyrimo rodikliuose nebuvo, taip pat ir tinkamo antibiotikų skyrimo srityje (koreguotas šansų santykis 0,86, 95 % PI 0,48–1,55). Antibiotikų skyrimo dažnio priemonė nepakeitė.

Pacientai skirtumo nepajuto. Tarp 826 pacientų, užpildžiusių pasitenkinimo apklausą, pasitenkinimas abiejose grupėse buvo iš esmės vienodas, kaip ir konsultacijų trukmė.

Išlaidos šiek tiek krypo žemyn. Koreguotoje analizėje su antibiotikais susijusios išlaidos DI grupėje buvo mažesnės – tikėtina dėl pigesnių vaistų pasirinkimo, o ne mažesnio receptų skaičiaus. Vienam pacientui tenkanti antibiotikų ekonomija atrodė didesnė nei vienam pacientui tenkanti priemonės naudojimo kaina. Tačiau autoriai tai įvardija kaip užuominą, ne galutinį rezultatą: visas sistemos nuosavybės ir eksploatavimo sąnaudas vertinanti ekonominė analizė į tyrimą neįėjo.

Pačių autorių vieno sakinio santrauka yra tiksliausia: šiose ribose LLM pagalba neparodė saugos problemų, bet per 14 dienų nesumažino gydymo nesėkmių, o jei nauda yra, ji greičiausiai nedidelė.

Kodėl „reikšmingo skirtumo nėra“ nereikia „priemonė neveikia“

Nulinę pirminę baigtį lengva perinterpretuoti į abi puses. Du dalykai neleidžia pasirinkti pernelyg paprastos istorijos.

Pirma, **tyrimas buvo suplanuotas aptikti didesnę poveikį, nei iš tikrųjų pamatė.** Sunkios nepalankios

baigtys pirminėje priežiūroje retos – čia apie 2 % – todėl atskirti mažą tikrą naudą nuo statistinio triukšmo reikia milžiniškos imties. Pačių autorių post hoc galios skaičiavimai rodo, kad tokio dydžio poveikiui, kokį jie stebėjo, aptikti reikėtų **gerokai didesnio tyrimo – daugiau kaip 100 000 pacientų eilės**. Statistiškai nereikšmingas rezultatas tokio dydžio tyrime neatmeta nedidelės tikros naudos; jis reiškia, kad šis tyrimas jos negalėjo patikimai atskirti.

Antra, **palyginimas iš dalies susiliejo**. Dėl konfigūracijos klaidos kai kurie kontrolinės grupės klinicistai trumpam gavo prieigą prie AI Consult, o viename tinkle dirbantys žmonės kalbasi tarpusavyje ir perneša įpročius per grupių ribą. Abu reiškiniai mažina skirtumą tarp grupių ir galimą tikrą poveikį stumia link nulio. Be to, pats tinklas jau dirbo gana aukštu standartu, tad priemonei liko mažiau erdvės pagerinimui.

Nė viena iš šių išlygų neišgelbsti „proveržio“ antraštės. Tačiau jos reiškia, kad teisingas skaitymas turi būti kalibruotas, o ne kategoriškai neigiamas: pagal sunkiausią ir sąžiningiausią baigtį priemonė per dvi savaites neparodė išmatuojamos naudos pacientams, nors dokumentaciją pagerino ir saugos signalo nekėlė.

Ko tai *neįrodo*

- Tyrimas **neparodo**, kad DI pagerino pacientų baigtis. Pagal pirminį 14 dienų rodiklį reikšmingos naudos nebuvo.
- Jis **neparodo**, kad DI bevertis. Taškinis įvertis krypto naudos link, dokumentacija pagerėjo, vaistų išlaidos turėjo mažėjimo tendenciją; nulinis rezultatas suderinamas su nedidele tikra nauda, kurios šis tyrimas buvo per mažas patvirtinti.
- Jis **neįrodo**, kad priemonė saugi retų pakenkimų atžvilgiu. Saugos signalo nebuvo, bet tyrimas nebuvo sukurtas sertifikuoti saugumą retoms sunkioms baigtims.
- Jis **nerodo**, kad „DI pranoksta gydytojus“ ar juos pakeičia. Tai sprendimų *palaikymas*; klinicistas išlaikė visą teisę pasiūlymą priimti ar atmesti.
- Rezultatai **nepersikelia automatiškai** į kitas aplinkas. Tyrimas vyko viename privačiame miesto tinkle Kenijoje; kaimo, priemiesčio ar didesnių pajamų sistemose poveikis gali būti kitoks į abi puses.
- Jis **neįrodo sutaupymo**. Išlaidų signalas įdomus, bet tai ne pilna ekonominė analizė.

Kiek tvirti yra įrodymai?

Pagrindiniam teiginiui – kad neįrodyta trumpalaikė paciento nauda – tyrimo dizainas yra stiprus, o išvada tinkamai kukli. Prospektyvus, iš anksto registruotas, klasterinis randomizuotas tyrimas su akiai vertinta paciento lygmens sudėtine baigtimi yra arti geriausių realaus pasaulio įrodymų, kuriuos galima surinkti tokiai priemonei. Jis daug informatyvesnis už etalono balą ar vinječių tyrimą.

Antriniam rezultatams – geresnei dokumentacijai, nepakitusiam vaistų skyrimui, panašiam pasitenkinimui, mažesnėms antibiotikų išlaidoms – įrodymai geri, tačiau juos reikia vertinti kaip

antrinius signalus, pažeidžiamus tos pačios grupių taršos ir vieno tinklo ribotumo.

Saugos duomenys ramina, bet yra riboti: signalo nerasta, tačiau tyrimas nebuvo skirtas retoms žaloms aptikti.

Naudingiausia pozicija nėra nei „veikia“, nei „nepavyko“. Tikslesnė: atsargus realaus pasaulio tyrimas parodė, kad ši DI priemonė nekėlė saugos signalo ir pagerino priežiūros procesą, bet per dvi savaites neįrodė pacientų baigties naudos; tokio dydžio naudai aptikti reikėtų daug didesnio tyrimo.

Kodėl tai svarbu

Medicininio DI diskusijai labai trūksta būtent tokių įrodymų. Yra tūkstančiai publikacijų, kuriose modeliai puikiai sprendžia egzaminus ir prilygsta kliniciams tvarkingose situacijose. Didelių, pragmatinių, randomizuotų tyrimų, matuojančių, ar tikri pacientai iš tiesų gauna naudos, yra labai mažai. Šis – vienas iš jų, ir jo rezultatas neblizga, bet svarbus: išlaikyti testą nėra tas pats, kas padėti pacientui.

Visa istorija slypi būtent šiame tarpsnyje. Priemonė gali būti realiai naudinga kliniciui – aiškesni įrašai, mažesnės vaistų sąskaitos, antra nuomonė – ir vis tiek per dvi savaites nepajudinti griežtos paciento baigties. Abu dalykai gali būti teisingi vienu metu, o brandi sveikatos sistema turi sugebėti juos laikyti kartu, o ne pasirinkti patogesnę.

Tyrimas taip pat naudingai perkelia įrodinėjimo našą. Jei bendrovė nori teigti, kad jos klinikinis DI gerina priežiūrą, aktualus įrodymas nėra reitingų lentelė. Tai tokio tipo tyrimas su pacientams svarbiomis baigtimis – ir, idealiu atveju, dar didesnis, nes sąžininga šio darbo pamoka yra ta, kad nedidelėms naudoms pamatyti reikia labai didelių imčių.

Trumpa santrauka

Pragmatiniame klasteriniame randomizuotame tyrime 16 Kenijos pirminės priežiūros įstaigų išbandyta generatyvinio DI sprendimų palaikymo priemonė AI Consult, integruota į clinical officers naudojamą elektroninį įrašą. Tarp 9 691 paciento ekspertų vertintas sudėtinis gydymo nesėkmės per 14 dienų rodiklis buvo 2,2 % su priemone ir 2,0 % be jos (koreguotas šansų santykis 0,77, 95 % PI 0,55–1,08, $P = 0,13$) – reikšmingo skirtumo nebuvo. Saugos signalo nenustatyta, klinikinė dokumentacija pagerėjo visose vertintose srityse, vaistų skyrimas ir pacientų pasitenkinimas nepasikeitė, o antibiotikų išlaidos buvo kiek mažesnės. Autoriai daro išvadą, kad šiose ribose priemonė atrodė saugi, tačiau nesumažino gydymo nesėkmių; jei nauda egzistuoja, ji tikriausiai nedidelė, o tokio dydžio poveikiui nustatyti reikėtų gerokai didesnio – maždaug 100 000 pacientų eilės – tyrimo. Rezultatas neparodo nei kad klinikinis DI gerina pacientų baigtis, nei kad jis nenaudingas. Jis parodo, kad realia mieste tinkla priemonė, kuri pagerino procesą ir nekėlė saugos signalo, per dvi savaites neparodė išmatuojamos naudos pacientams.

Be pagražinimų

Ką rodo tyrimas: realiame randomizuotame tyrime LLM pagrįstos sprendimų palaikymo priemonės pridėjimas prie pirminės priežiūros įrašų nekėlė saugos signalo, pagerino dokumentacijos kokybę, nepakeitė vaistų skyrimo ir reikšmingai nesumažino griežto 14 dienų gydymo nesėkmės rodiklio.

Kas tikėtina, bet neįrodyta: kad priemonė duoda nedidelį tikrą gydymo nesėkmių sumažėjimą, kurio šis tyrimas per mažas aptikti; kad suskaičiavus visas sąnaudas ji taupo pinigus; kad geresnė dokumentacija vėliau virsta geresne priežiūra.

Ko tyrimas nerodo: kad klinikinis DI gerina pacientų baigtis; kad priemonė saugi retų reiškinių atžvilgiu; kad ji pakeičia ar pranoksta klinikistus; kad rezultatai automatiškai galioja kaimo ar didesnių pajamų aplinkoms; kad sutaupymas jau įrodytas.

Pagrindiniai ribotumai: tyrimas suplanuotas didesniame poveikiui nei stebėtas, o retos baigtys reikalauja didžiulių imčių; konfigūracijos klaida trumpam užteršė kontrolinę grupę, o bendro tinklo darbo įpročiai taip pat mažino grupių skirtumą, abu efektai stumdami rezultatą link nulio; vienas privatus miesto tinklas su jau gana aukštais standartais; nebuvo iš anksto numatyto ne mažesnio veiksmingumo ar formalaus saugos dizaino; stebėjimas truko 14 dienų.

Kiek pasitikėti turėtų bendrasis skaitytojas? Tvirtai – kad šiame tyrime saugos signalo nebuvo ir dokumentacija pagerėjo. Tvirtai – kad čia nebuvo parodytas išmatuojamas 14 dienų pacientų baigčių pagerėjimas. Menkai – bet koku kategorišku teiginiu, kad priemonė kaip paciento naudos intervencija „veikia“ arba „neveikia“; šis klausimas iš tiesų lieka neišspręstas ir reikalauja daug didesnio tyrimo. Tinkama pozicija: kruopštus realaus pasaulio rezultatas apie priemonę, kuri pagerino priežiūros procesą ir nekėlė saugos signalo, o ne nuosprendis, kad DI išgelbės ar sugadins pirminę priežiūrą.

Šaltiniai

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/>

EurekaAlert / PATH press release on the trial (2026)

<https://www.eurekaalert.org/news-releases/1133583>