



The CLEAN PAPER

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

Clinical AI kan tó dà bí safe, tó sì improve paperwork — sùgbọn kò improve patient outcomes

Pragmatic cluster-randomized trial nínú 16 Kenyan primary care facilities test generative-AI decision-support tool tí a fi kún record clinical officers. Nínú 9,691 patients, strict 14-day expert-adjudicated composite of treatment failure jé 2.2% pèlú tool versus 2.0% láisí rẹ (adjusted odds ratio 0.77, 95% CI 0.55-1.08, $P = 0.13$) — no significant difference. Tool kò fi safety signal hàn, ó improve documentation kojá all rated domains, prescribing àti satisfaction unchanged, antibiotic costs sì trend lower. Èyí kì í ẹ failed study; ó jé rare, honest one — measured on patients, not benchmarks — tó fi hàn pé tool tí looked safe, tó help process, kò demonstrably help patients nínú two weeks, benefit kankan sì yòò nilo much larger trial láti detect.

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

Safe àti helpful kì í dọgba pèlú effective

Òpò àkọlé nípa ìṣẹ̀gùn AI ni a kọ láti wrong kind of ìwádíí. àwòṣe kan score dáadàa lórí exam ibéèrè, tàbí ṣẹ̀gun doctors lórí curated vignettes, itàn sì kọ ara rẹ: machine ti ready fún clinic.

ìdánwò yíi ṣe ohun tó nira, tó sì ṣòwọ̀n. Ó fi generative-AI decision-support irinṣẹ́ sínú gidi primary care, nínú gidi facilities, pèlú gidi aláìsàn, ó sì béèrè ibéèrè tó matter gidi: ṣe aláìsàn ṣe dáadàa sí i?

Honest answer ni no — kò measurable, kì í ṣe nínú 14 days. irinṣẹ́ náà kò fi ààbò àmì hàn. Ó mú sunwọ̀n sí i quality ìṣẹ̀gùn documentation. Ó lè ti lower diẹ̀ drug costs. Ṣùgbọ̀n kò significantly dín kù ìtoju ìkùná, àwọ̀n olùkòwé sì ṣọ̀ra láti sọ pé àhàfàní kankan, tí ó bá wà, probably modest.

Èyí kì í ṣe ìkùná of ìwádíí. Èyí ni ìwádíí tí ń ṣiṣẹ́. Irú ẹ̀rí responsible nípa AI nínú medicine ni èyí, nígbà tí a bá measure lórí aláìsàn dípò benchmarks.

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

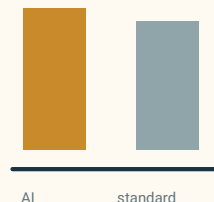
Process signal, not outcome proof.



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.

irinṣẹ́ náà kò fi ààbò àmì hàn, ó sì mú sunwọ̀n sí i ìṣẹ̀gùn documentation, ṣùgbọ̀n prespecified 14-day aláìsàn èsì kò significantly mú sunwọ̀n sí i. Process help kì í dọgba pèlú proven aláìsàn àhàfàní.

The Clean Paper CC BY 4.0

Ohun tí àwọ̀n olùkòwé ṣe

ẹgbẹ́ náà run pragmatic, cluster-randomized ìdánwò kojá **16 primary care facilities** tí private health network (Penda Health) ń operate ní Nairobi àti Kiambu counties, Kenya. Care nínú facilities wọ̀nyí jẹ́ mostly láti ọwọ́

ìṣẹ̀gùn officers — mid-level practitioners pẹ̀lú three-year diploma — tí wọn kò sáà ní easy access sí senior consultation.

Unit of randomization ni clinician, kì í ṣe aláìsàn. **103 ìṣẹ̀gùn officers** ni wọn randomize: 52 sí intervention arm àti 51 sí control arm. Arms méjèèjì lo kan nàà cloud-based electronic ìṣẹ̀gùn àkòsílẹ̀. Intervention arm ní afikun “**AI Consult**” (ẹ̀dà 2.0), decision-support irinṣẹ̀ tí a kọ lórí **OpenAI’s GPT-4o** large ẹ̀dè àwòṣe, tí a embed sínú àkòsílẹ̀ nàà. Ó ka information tí clinician document, ó sì lè flag ẹ́ẹ̀ ṣe issues pẹ̀lú ìdánimọ̀ àrùn tàbí ìtoju plan. Clinicians pa full autonomy mó: wọn lè accept, modify tàbí ignore suggestions rẹ̀.

Model wo ni, kí ló sì dé tí specifics matter?

irinṣẹ̀ nàà jẹ **AI Consult 2.0**, tó run **OpenAI’s GPT-4o** (May 2025 release), accessed nipasẹ OpenAI commercial API labẹ enterprise licence, pẹ̀lú low-randomness settings (temperature 0.1). Ó wà nínú bespoke electronic àkòsílẹ̀ (EasyClinic’s EMR), ètò prompts sì steer rẹ̀ láti align pẹ̀lú Kenyan national ìtoju guidelines; àwọn olùkòwé publish full instruction prompt.

Kí ló dé tá a fi sọ gbogbo èyí? Nítorí èsì jẹ nípa *specific ètò kan* — àwòṣe ẹ̀dà kan, prompt kan, àkòsílẹ̀ kan, setting kan — kì í ṣe “LLMs in medicine” ní gbogbogbo. Àwọn olùkòwé ṣe kan nàà kókó: wọn pe àwárí wọn ní *temporal benchmark rather than fixed ìṣíro àfojúsùn of capability*. Newer àwòṣe, yàtò prompt, tàbí clinic tí digitization rẹ̀ kéré lè gbogbo wọn yí èsì padà.

Nípa independence: OpenAI nigbamii fún in-kind ṣe àtílẹ̀yìn fún (cloud-compute credits àti technical guidance lórí API), ṣùgbọ́n àwọn olùkòwé sọ pé decision láti lo OpenAI ṣẹ̀lẹ̀ *ṣáájú* offer yẹn, àti pé OpenAI kò ní role nínú ìdánwò àpẹ̀rẹ̀, dátà collection, ìtupalẹ̀ tàbí decision to publish.

Láàárín 22 April àti 16 July 2025, **9,691 aláìsàn** ni wọn enroll. **Primary èsì jẹ deliberately patient-centered àti strict:** expert-adjudicated *composite of ìtoju ìkùnà* nínú 14 days léyìn visit — panel clinicians, blind sí iwádíí arm, judge bóyá aláìsàn kọ́okan ní bad èsì bí unresolved tàbí worsening illness. ìdánwò nàà ni a register *ṣáájú* (Pan-African Clinical Trials Registry 202502499779176).

Design choice yí ni kókó. Ó rọ̀rùn láti fi hàn pé AI irinṣẹ̀ yí ohun tí clinician kọ́ sílẹ̀ padà. Ó nira púpọ̀, ó sì meaningful sí i, láti fi hàn pé ó yí ohun tí ó ṣẹ̀lẹ̀ sí aláìsàn padà.

Ohun tí wọn rí

Primary èsì kò mú sunwọn sí i. ìtoju ìkùnà ṣẹ̀lẹ̀ nínú 102 of 4,693 aláìsàn (2.2%) nínú AI arm àti 94 of 4,654 (2.0%) nínú control arm. Crude percentages díẹ̀ ga pẹ̀lú AI, ṣùgbọ́n léyìn adjustment kókó ìṣíro àfojúsùn lean toward ànfaàní: **adjusted odds ratio 0.77 (95% ìgbẹ̀kẹ̀lé interval 0.55 to 1.08, P = 0.13)** — not statistically significant. Flip láàárín raw àti adjusted numbers kì í ṣe arithmetic slip; adjustment accounts fún differences láàárín clinician clusters. Ní ọ̀nà méjèèjì, ìgbẹ̀kẹ̀lé interval comfortably include “no ipa,” nítorí nàà a kò lè claim ànfaàní, absolute ipa sì tiny.

Fún plain-English way láti ka odds ratios, ìgbẹ̀kẹ̀lé intervals àti P values papọ̀, wo guide to ìṣẹ̀gùn èsì.

Kò sí ààbò àmì — lábẹ̀ ààlà. Kò sí serious adverse ìṣẹ̀lẹ̀ tí a judge gégé bí related sí irinṣẹ̀, independent review sì kò rí ààbò àmì. Àwọn olùkòwé honest nípa ceiling reassurance yí: ìdánwò kò powered láti detect rare severe harms,

kò sí ní prespecified noninferiority tàbí formal ààbò àgbékalè, nítorí náà kò lè *fi èrí múlè* ààbò fún uncommon ìṣẹ̀lẹ̀.

Documentation dára sí i. Nínú encounters 2,000 tí blinded experts review, clinicians tí wọn lo AI Consult produce better ìṣẹ̀gùn documentation kojá gbogbo rated domains — recorded idánimọ̀ àrùn, itoju plan àti overall completeness.

Prescribing fẹ̀rẹ̀ má yí. Kò sí significant difference nínú prescribing, pẹ̀lú correct antibiotic lò (adjusted odds ratio 0.86, 95% CI 0.48 to 1.55). irinṣẹ̀ kò yí antibiotic prescribing rates padà.

alàìsàn kò notice difference. Nínú alàìsàn 826 tí wọn complete satisfaction survey, satisfaction fẹ̀rẹ̀ identical láàárín arms, consultation times sì similar.

Costs tọ̀ka diẹ̀ sí isalẹ̀. Nínú adjusted itupalẹ̀, antibiotic-related costs kéré nínú AI arm — plausibly nítorí cheaper choices dípò fewer prescriptions — per-patient antibiotic saving sì dà bí ẹ̀ni pé ó ju per-patient cost ti running irinṣẹ̀. Àwọn olùkòwé flag èyí gégé bí suggestive, kì í ṣe settled: full total-cost-of-ownership accounting wà níta idánwò.

One-line summary àwọn olùkòwé fúnra wọn ni cleanest èdà: LLM assistance láiléwu lábé ààlà wònyí, ṣùgbọ̀n kò dín kù itoju ikùnà nínú 14 days, àhàfàní kankan probably modest.

Kí ló dé tí “no significant difference” kì í ṣe “it doesn’t ìṣẹ̀”

Null primary èsì rọ̀rùn láti over-read ní itòṣọ̀nà méjì. Nìkan méjì dá simple itàn dúró.

Àkọ́kọ́, **idánwò náà built láti catch ipa tó tóbi ju ohun tí ó rí lọ.** Serious bad èsì nínú primary care rare — tó 2% níbí — nítorí náà distinguishing small gidi àhàfàní láti ariwo nílò enormous numbers. Post-hoc power calculations àwọn olùkòwé daba pé láti detect ipa iwọn tí wọn observe yóò nílò **idánwò tó tóbi gan-an, order of sí i than 100,000 alàìsàn.** Non-significant èsì nínú iwádíí iwọn yí kò òfin out small gidi àhàfàní; ó tùmọ̀ sí pé iwádíí yí kò lè resolve rẹ̀.

Elẹ̀kẹ̀jẹ̀, **ifiwéra náà blurred diẹ̀.** Configuration àṣìṣe kan fún brief access sí AI Consult fún diẹ̀ control-arm clinicians, clinicians nínú shared network sì ní bá ara wọn sọ̀rọ̀, wọn sì carry habits kojá ààlà. Effects méjèjèjì ní mú arms méjèjèjì jọra sí i, pushing any gidi difference toward zero. Lóri rẹ̀, host network ti already run sí relatively high standards, tó fi room kéré sílẹ̀ fún irinṣẹ̀ láti fi hàn improvement.

Kò sí èyí tó rescue “breakthrough” àkọ́lẹ̀. Ṣùgbọ̀n ó tùmọ̀ sí pé correct reading jẹ̀ calibrated, kì í ṣe deflationary: lóri hardest, most honest endpoint, irinṣẹ̀ yí kò demonstrably help alàìsàn nínú méjì weeks — nígbà tí ó measurably help record-keeping, ó sì looked láiléwu.

Ohun tí èyí kò fi èrí múlè

- Kò fi hàn pé AI mú sunwọn sí i alàìsàn èsì. Lóri primary 14-day endpoint, kò sí significant àhàfàní.
- Kò fi hàn pé AI useless. Point ìṣìrò àfọ̀júsùn favored rẹ̀, documentation mú sunwọn sí i jálẹ̀, drug costs trend down; null èsì consistent pẹ̀lú small gidi àhàfàní tí idánwò kere ju láti confirm.

- Kò fi ẹrì múlẹ̀ irinṣẹ̀ láilẹ̀wu fún rare harms. Kò fi ààbò àmì hàn, ṣùgbọ̀n idánwò kò powered tàbí designed láti certify ààbò fún uncommon severe ìṣẹ̀lẹ̀.
- Kò fi hàn pé “AI beats doctors” tàbí replaces clinicians. Èyí jẹ́ decision *ṣe àtìlẹ́yìn fún*; clinician pa full authority mó láti accept tàbí reject.
- Kò generalize automatically. idánwò run nínú single private urban network ní Kenya; rural, periurban àti higher-income settings lè differ ní ìtọ̀sọ̀nà kankan.
- Kò establish cost savings. Cost àmì suggestive, kì í ṣe completed economic evaluation.

Báwo ni ẹrì ṣe lágbára tó?

Fún central claim — no proven reduction nínú short-term aláìsàn harm — ẹrì lágbára *gégé bí àpẹrẹ*, appropriately humble *gégé bí conclusion*. Prospective, pre-registered, cluster-randomized idánwò pèlú blinded patient-level composite èsì sún mó best ayé gidi ẹrì tí a lè gather fún irinṣẹ̀ bááyí. Ó informative púpọ̀ ju benchmark score tàbí vignette iwádíí lọ.

Fún secondary àwárí — better documentation, unchanged prescribing, similar satisfaction, lower antibiotic costs — ẹrì good ṣùgbọ̀n yẹ kí a ka wọn gégé bí secondary: supportive àmì, kì í ṣe àkọlé, wọn sì vulnerable sí contamination àti single-network ààlà.

Fún ààbò, ẹrì reassuring ṣùgbọ̀n bounded: no àmì rí, ṣùgbọ̀n kì í ṣe iwádíí built láti rí rare harms.

Most useful stance kì í ṣe “it works” tàbí “it failed.” Ó jẹ́: careful ayé gidi idánwò rí pé AI irinṣẹ̀ yí kò raise ààbò àmì, ó mú sunwọn sí i ilànà of care, ṣùgbọ̀n kò demonstrate patient-outcome ànfààní nínú méjì weeks — àti detecting such ànfààní yóò nilo iwádíí tó tóbi jù lọ.

Kí nìdí tí ó fi ṣe pàtàkì?

Debate nípa ìṣẹ̀gùn AI ebi ní pa fún irú ẹrì yí. Thousands of ìwé iwádíí wà tí àwòṣe ní ace exams, match clinicians lórí tidy ọ̀ràn. Large pragmatic randomized idánwò díẹ̀ ní wà tí ní measure bóyá gidi aláìsàn better off. Èyí jẹ́ ọ̀kan wọn, ó sì land lórí unglamorous truth: passing idánwò kì í dọ̀gba pèlú helping aláìsàn.

Gap yẹn ní gbogbo ìtàn. irinṣẹ̀ kan lè genuinely useful sí clinicians — clearer notes, lower drug bills, second pair of eyes — ṣùgbọ̀n sì má move hard aláìsàn èsì nínú fortnight. Facts méjèjèjì lè true ní akoko kan náà, mature health ètò gbọ̀dọ̀ hold wọn papọ̀ dípò pick convenient ọ̀kan.

Ó tún reset burden of proof lònà useful. Tí company bá fẹ́ claim pé ìṣẹ̀gùn AI rẹ̀ mú sunwọn sí i care, relevant ẹrì kì í ṣe leaderboard. Ó jẹ́ idánwò bí yí, lórí èsì tí matter sí aláìsàn — ideally bigger ọ̀kan, nítorí honest lesson níbi ní pé modest ànfààní nílò large iwádíí láti rí.

Àkótán kedere

Pragmatic cluster-randomized ìdánwò nínú 16 Kenyan primary care facilities ìdánwò generative-AI decision-support irinṣẹ́ (“AI Consult”) tí a fi kún electronic àkọsílẹ̀ tí ìṣẹ̀gùn officers ní lo. Nínú 9,691 aláìsàn, expert-adjudicated composite ìtoju ìkùnà nínú 14 days jẹ 2.2% pẹ̀lú irinṣẹ́ versus 2.0% láìsí rẹ̀ (adjusted odds ratio 0.77, 95% CI 0.55–1.08, $P = 0.13$) — no significant difference. irinṣẹ́ kò fi ààbò àmì hàn, ó mú sunwọn sí i ìṣẹ̀gùn documentation kọ́já gbogbo rated domains, kò yí prescribing padà, aláìsàn satisfaction unchanged, ó sì associated pẹ̀lú somewhat lower antibiotic costs. Àwọn olùkòwé conclude pé ó láiléwu lábẹ̀ ààlà wọ̀nyí, sùgbọ̀n kò dín kù ìtoju ìkùnà, ànfààní kankan probably modest; detecting ipa iwọn tí wọ̀n observe yóò nilo ìdánwò tó tóbi gan-an (order of 100,000 aláìsàn). èsì kò fi hàn pé ìṣẹ̀gùn AI mú sunwọn sí i aláìsàn èsì, kò sì fi hàn pé useless — ó fi hàn pé irinṣẹ́ tí ó looked láiléwu, tí ó help ilàà, kò demonstrably help aláìsàn nínú méjì weeks nínú urban private network kan.

Àyèwò láìsí àṣejù

Ohun tí iwé iwádíí fi hàn: Nínú ayé gidi randomized ìdánwò, adding LLM-based decision-support irinṣẹ́ sí primary care àkọsílẹ̀ kò raise ààbò àmì, mú sunwọn sí i ìṣẹ̀gùn documentation quality, kò yí prescribing padà, kò sì significantly dín kù strict 14-day aláìsàn composite of ìtoju ìkùnà.

Ohun tó ṣeé gbà sùgbọ̀n tí a kò fi ẹrì múlẹ̀: Pé irinṣẹ́ produce small gidi reduction nínú ìtoju ìkùnà tó kéré ju ìdánwò yí láti detect; pé ó save money léyìn full costs; pé better documentation eventually translate sí better care.

Ohun tí kò fi hàn: Pé ìṣẹ̀gùn AI mú sunwọn sí i aláìsàn èsì; pé ó unsafe fún rare ìṣẹ̀lẹ̀; pé ó replace tàbí outperform clinicians; pé èsì transfer sí rural tàbí higher-income settings; pé cost saving established.

Main limitations: Powered fún larger ipa ju observed lọ (rare èsì nílò huge àpẹẹrẹ); configuration àṣiṣe contaminate control arm, shared-network habits sì blur ifiwéra, méjèèjì bias toward no difference; single private urban network pẹ̀lú already-high standards; no prespecified noninferiority tàbí ààbò àgbékalẹ̀; short 14-day horizon.

Confidence wo ni gbogbogbò reader yẹ kí ó ní? Gíga pé irinṣẹ́ kò fi ààbò àmì hàn nínú ìdánwò yí, ó sì mú sunwọn sí i documentation. Gíga pé kò demonstrably mú sunwọn sí i 14-day aláìsàn èsì nìbí. Kéré fún claim kankan pé ó “works” tàbí “fails” gégé bí patient-benefit intervention — ìbèèrè nàà genuinely unresolved, ó sì nílò much larger iwádíí. Ìpò tó yẹ: careful ayé gidi èsì nípa irinṣẹ́ tí kò raise ààbò àmì, tó mú sunwọn sí i ilàà of care, kì í ṣe verdict pé AI transforms tàbí wrecks primary care.

Àwọn orísun

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>