The Clean Paper

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

Təhlükəsiz görünən və sənədləşdirməni yaxşılaşdıran klinik AI — amma pasiyent nəticələrini yaxşılaşdırmadı

Keniyada 16 primary care facility-də pragmatic, cluster-randomized trial clinical officers-un record-una əlavə edilən generative-AI decision-support alətini yoxladı. 9,691 pasiyent arasında sərt 14-day expert-adjudicated treatment-failure composite alətlə 2.2%, alətsiz 2.0% idi (adjusted odds ratio 0.77, 95% CI 0.55–1.08, P = 0.13) — əhəmiyyətli fərq yoxdur. Alət safety signal göstərmədi, bütün sahələrdə documentation-u yaxşılaşdırdı, prescribing və satisfaction-u dəyişmədi və aşağı antibiotic costs istiqaməti göstərdi. Bu uğursuz tədqiqat deyil; benchmarkda yox, pasiyentlərdə ölçülən nadir dürüst tədqiqatdır — təhlükəsiz görünən və prosesi yaxşılaşdıran alətin iki həftədə pasiyentlərə nümayiş etdirilən fayda vermədiyi və hər hansı belə faydanın aşkarlanması üçün çox daha böyük trial lazım olduğu glamoursuz nəticəsinə gəlir.

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

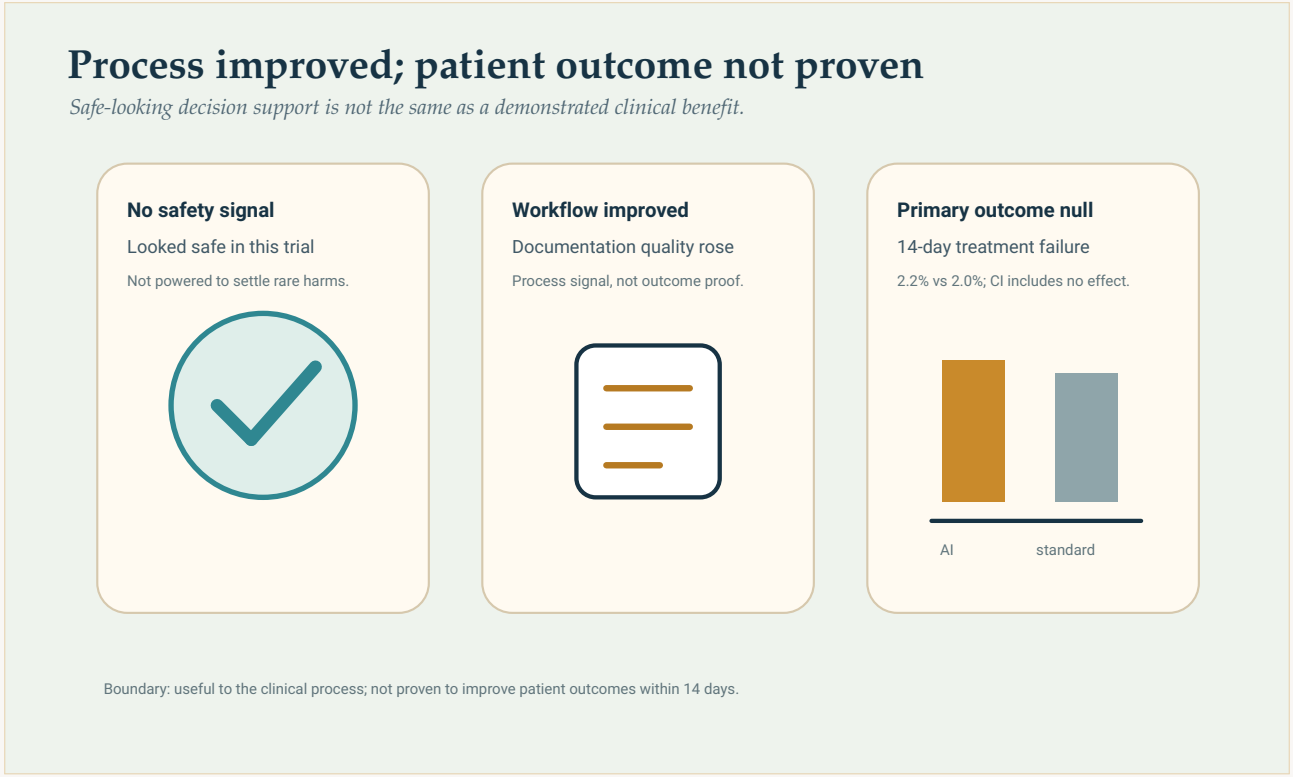
Təhlükəsiz və faydalı görünmək effektiv olmaqla eyni deyil

Tibbi AI haqqında başlıqların çoxu yanlış tip tədqiqatdan qurulur. Model imtahan suallarında yüksək bal alır və ya seçilmiş klinik ssenarilərdə həkimləri keçir, hekayə özü yazılır: maşın klinikaya hazırdır.

Bu sınaq daha çətin və nadir bir iş gördü. Generativ-AI qərar dəstəyi alətini real primary care-a, real müəssisələrə, real pasiyentlərə yerləşdirdi və əsl vacib sualı verdi: pasiyentlərin vəziyyəti yaxşılaşdımı?

Dürüst cavab xeyrdir — ölçülə bilən şəkildə yox, 14 gün ərzində yox. Alət təhlükəsizlik siqnalı göstərmədi. Klinik sənədləşdirmənin keyfiyyətini yaxşılaşdırdı. Hətta bəzi dərman xərclərini azaltmış ola bilər. Amma treatment failure-ları əhəmiyyətli dərəcədə azaltmadı və müəlliflər ehtiyatla deyir ki, fayda varsa, yəqin mülayimdir.

Bu, tədqiqatın uğursuzluğu deyil. Tədqiqatın işləməsidir. Tibbdə AI haqqında məsuliyyətli sübut benchmarklarda yox, pasiyentlərdə ölçüləndə belə görünür.



Alət təhlükəsizlik siqnalı göstərmədi və klinik sənədləşdirməni yaxşılaşdırdı, amma əvvəlcədən müəyyən edilmiş 14 günlük pasiyent nəticəsi əhəmiyyətli dərəcədə yaxşılaşmadı. Prosesə kömək sübut edilmiş pasiyent faydası ilə eyni deyil.

Müəlliflər nə etdilər

Komanda Keniyanın Nairobi və Kiambu bölgələrində özəl səhiyyə şəbəkəsi (Penda Health) tərəfindən işlədilən **16 primary care facility** üzrə pragmatic, cluster-randomized trial apardı. Bu müəssisələrdə xidməti əsasən **clinical officers** — üçillik diplomlu orta-səviyyə praktikantlar — göstərir və onların senior konsultasiyaya asan çıxışı çox vaxt yoxdur.

Randomizasiya vahidi pasiyent yox, klinisist idi. **103 clinical officer** randomizə edildi: 52 intervention arm-a, 51 control arm-a. Hər iki qrup eyni cloud-based electronic medical record istifadə edirdi. Intervention arm əlavə olaraq həmin record daxilində qurulmuş, **OpenAI GPT-4o** large language model əsasında **“AI Consult”** (version 2.0) qərar-dəstək alətinə sahib idi. Alət klinisistin sənədləşdirdiyi məlumatı oxuyur və diaqnoz və ya treatment plan ilə bağlı mümkün problemləri işarə edə bilirdi. Klinisistlərin tam avtonomiyası qalırdı: təklifləri qəbul, dəyişdirə və ya ignor edə bilərdilər.

> **Hansı model idi və niyə detallar vacibdir**
>
> Alət **AI Consult 2.0** idi, **OpenAI GPT-4o**-nun May 2025 versiyasını işlədirdi; OpenAI-nin commercial API-si vasitəsilə enterprise licence altında və aşağı-randomness setting-də (temperature 0.1) istifadə olunurdu. EasyClinic-in bespoke EMR-i daxilində idi və Kenyan national treatment guidelines ilə uyğunlaşmaq üçün yazılmış system prompts tərəfindən yönləndirilirdi; müəlliflər tam instruction prompt-u dərc ediblər.
> Niyə bunu bu qədər dəqiq deyirik? Çünki nəticə ümumən “tibbdə LLM-lər” barədə yox, *bir konkret sistem* — bir model versiyası, bir prompt, bir record, bir setting — haqqındadır. Müəlliflər də eyni məqamı vurğulayırlar: tapıntını sabit qabiliyyət qiyməti yox, *temporal benchmark* adlandırırlar. Daha yeni model, fərqli prompt və ya daha az rəqəmsallaşmış klinika nəticəni dəyişə bilər.
> Müstəqillik barədə: OpenAI sonradan in-kind support (cloud-compute credits və API istifadəsi üçün texniki guidance) verdi, amma müəlliflər OpenAI-dan istifadə qərarının bu təklifdən *əvvəl* verildiyini və OpenAI-nin trial dizaynı, data collection, analysis və publication qərarında rol oynamadığını bildirirlər.

22 April–16 July 2025 arasında **9,691 pasiyent** daxil edildi. **Primary outcome qəsdən pasiyent-mərkəzli və sərt idi**: ziyarətdən sonrakı 14 gündə expert-adjudicated *composite of treatment failure* — study arm-dan xəbərsiz klinisist paneli hər pasiyentin həll olunmayan və ya pisləşən xəstəlik kimi pis nəticə yaşayıb-yaşamadığını qiymətləndirdi. Trial əvvəlcədən qeydiyyatdan keçirilmişdi (Pan-African Clinical Trials Registry 202502499779176).

Dizayn seçimi məhz əsas məqamdır. AI alətinin klinisistin yazdıqlarını dəyişdirdiyini göstərmək asandır. Pasiyentin başına gələnləri dəyişdirdiyini göstərmək daha çətin və daha mənalıdır.

Nə tapdılar

Primary outcome yaxşılaşmadı. Treatment failure AI arm-da 4,693 pasiyentdən 102-də (2.2%), control arm-da 4,654-dən 94-də (2.0%) baş verdi. Xam faizlər AI ilə azacıq yüksək idi, amma adjustment-dan sonra point estimate fayda istiqamətinə döndü: **adjusted odds ratio 0.77 (95% confidence interval 0.55–1.08, P = 0.13)** — statistik əhəmiyyətli deyil. Xam və adjusted rəqəmlərin istiqamətinin dəyişməsi hesab səhvi deyil; adjustment clinician clusters arasındakı fərqləri nəzərə alır. Hər iki halda confidence interval rahatlıqla “no effect”i əhatə edir, deməli fayda iddia edilə bilməz və absolute effect çox kiçik idi.

Odds ratios, confidence intervals və P values-i birlikdə sadə dillə oxumaq üçün clinical results bələdçisinə baxın.

Təhlükəsizlik siqnalı yox idi — müəyyən sərhədlərdə. Alətlə əlaqəli hesab edilən serious adverse event olmadı və müstəqil review təhlükəsizlik siqnalı tapmadı. Müəlliflər bu rahatlığın tavanını dürüst deyirlər: trial nadir ağır zərərləri aşkarlamaq üçün powered deyildi və əvvəlcədən müəyyən edilmiş noninferiority və ya formal safety framework yox idi, buna görə nadir hadisələr üçün təhlükəsizliyi *sübut* edə bilməz.

Sənədləşdirmə yaxşılaşdı. Kor ekspertlər tərəfindən baxılan 2,000 encounter arasında AI Consult istifadə edən klinisistlər bütün qiymətləndirilən sahələrdə — yazılmış diaqnoz, treatment plan və ümumi completeness — daha yaxşı klinik sənədləşdirmə verdi.

Prescribing demək olar dəyişmədi. Correct antibiotic use daxil olmaqla prescribing-də statistik əhəmiyyətli fərq yox idi (adjusted odds ratio 0.86, 95% CI 0.48–1.55). Alət antibiotic prescribing rate-i dəyişmədi.

Pasiyentlər fərq hiss etmədi. Satisfaction survey dolduran 826 pasiyent arasında məmnunluq qruplarda demək olar eyni idi, consultation times də oxşar idi.

Xərclər bir qədər aşağı istiqamət göstərdi. Adjusted analysis-də antibiotic-related costs AI arm-da aşağı idi — yəqin daha az prescription-dan yox, daha ucuz seçimlərdən — və pasiyent başına antibiotic qənaəti alətin pasiyent başına işləmə xərcini aşırdı. Müəlliflər bunu settled nəticə yox, suggestive siqnal kimi göstərirlər: full total-cost-of-ownership trial xaricində idi.

Müəlliflərin öz bircümləlik xülasəsi ən təmiz versiyadır: LLM assistance bu məhdudiyyətlər daxilində təhlükəsiz idi, amma 14 gün ərzində treatment failure azaltmadı və hər hansı fayda yəqin mülayimdir.

Niyə “əhəmiyyətli fərq yoxdur” “işləmir” demək deyil

Null primary outcome-u hər iki istiqamətdə həddindən artıq oxumaq asandır. İki şey sadə hekayəyə mane olur.

Birincisi, **trial tapdığından daha böyük effekti tutmaq üçün qurulmuşdu.** Primary care-

da ciddi pis nəticələr nadirdir — burada təxminən 2% — buna görə kiçik real faydanı səs-küydən ayırmaq nəhəng saylar tələb edir. Müəlliflərin post-hoc power calculations-ı müşahidə etdikləri ölçüdə effekti aşkar etmək üçün **çox daha böyük, 100,000-dən çox pasiyent səviyyəsində trial** lazım ola biləcəyini göstərir. Bu ölçüdə tədqiqatda non-significant nəticə kiçik real faydanı istisna etmir; sadəcə bu tədqiqat onu ayırd edə bilmədi.

İkincisi, **müqayisə qismən bulanmışdı.** Configuration error qısa müddət bəzi control-arm klinisistlərinə AI Consult girişi verdi, üstəlik eyni şəbəkədə klinisistlər bir-biri ilə danışır və vərdişləri qruplar arasına daşıyır. Hər iki effekt qrupları bir-birinə daha oxşar göstərir və real fərqi sıfıra doğru çəkir. Bundan əlavə host network onsuz da nisbətən yüksək standartla işləyirdi, bu da alətin yaxşılaşma göstərməsi üçün az yer saxlayır.

Bunların heç biri “breakthrough” başlığını xilas etmir. Amma düzgün oxunuşun dismissive yox, calibrated olması deməkdir: ən çətin və dürüst endpoint-də bu alət iki həftədə pasiyentlərə nümayiş etdirilə bilən fayda vermədi — eyni zamanda record-keeping-i ölçülə bilən şəkildə yaxşılaşdırdı və təhlükəsiz göründü.

Bu nəyi *sübut etmir*

- AI-nın pasiyent nəticələrini yaxşılaşdırdığını **göstərmir**. Primary 14-day endpoint-də əhəmiyyətli fayda yox idi.
- AI-nın faydasız olduğunu **göstərmir**. Point estimate onun xeyrinə idi, documentation bütün sahələrdə yaxşılaşdı və drug costs aşağı istiqamət göstərdi; null nəticə trial-ın təsdiqləmək üçün kiçik olduğu real kiçik fayda ilə uyğundur.
- Nadir zərərlər üçün alətin təhlükəsiz olduğunu **sübut etmir**. Safety signal yox idi, amma nadir severe events üçün safety-ni certify etmək üçün powered və ya designed deyildi.
- “AI doctors-u keçir” və ya clinicians-i əvəz edir nəticəsini **göstərmir**. Bu decision *support*-dur; clinician qəbul və ya rədd etmək üçün tam səlahiyyəti saxladı.
- Avtomatik **generalize etmir**. Trial Keniyada bir özəl urban network-də aparılıb; rural, periurban və yüksək-gəlirli settings hər iki istiqamətdə fərqli ola bilər.
- Cost savings-i **müəyyən etmir**. Cost signal suggestive-dir, completed economic evaluation deyil.

Sübut nə qədər güclüdür?

Mərkəzi iddia — qısa-müddətli pasiyent zərərində sübut edilmiş azalma yoxdur — üçün sübut dizayn baxımından güclü, nəticə baxımından isə düzgün dərəcədə təvazökardır. Prospective, pre-registered, cluster-randomized trial və kor, patient-level composite outcome belə alət üçün toplanacaq ən yaxşı real-world evidence növlərindən biridir. Benchmark balından və vignette study-dən çox daha məlumatlıdır.

Secondary findings — daha yaxşı documentation, dəyişməyən prescribing, oxşar satisfaction,

aşağı antibiotic costs — üçün sübut yaxşıdır, amma secondary kimi oxunmalıdır: headline yox, supportive signals və eyni contamination/single-network limitlərinə həssas.

Safety üçün sübut reassuring, amma bounded-dir: siqnal tapılmadı, amma nadir zərərləri tapmaq üçün qurulmuş tədqiqat deyil.

Ən faydalı mövqe nə "işləyir", nə "uğursuz oldu"dur. Budur: diqqətli real-world trial bu AI alətinin safety signal yaratmadığını və care process-i yaxşılaşdırdığını, amma iki həftədə patient-outcome benefit nümayiş etdirmədiyini göstərdi — hər hansı belə faydanı tapmaq üçün daha böyük tədqiqat lazım olacaq.

Niyə vacibdir

Medical AI mübahisəsi məhz belə sübuta acdır. Modellərin imtahanları yüksək nəticə ilə verdiyini və səliqəli hallarda clinician-lərlə uyğunlaşdığını göstərən minlərlə məqalə var. Real pasiyentlərin daha yaxşı olub-olmadığını ölçən böyük, pragmatic, randomized trial isə çox azdır. Bu onlardan biridir və glamoursuz həqiqətə gəlir: testi keçmək pasiyentə kömək etməklə eyni deyil.

Bütün hekayə bu boşluqdur. Alət clinician-lər üçün həqiqətən faydalı ola bilər — daha aydın qeydlər, aşağı drug bills, ikinci cüt göz — və yenə də iki həftədə sərt patient outcome-u dəyişdirməyə bilər. Hər iki fakt eyni vaxtda doğru ola bilər və yetkin səhiyyə sistemi rahat olanını seçmək əvəzinə ikisini birlikdə saxlamalıdır.

Bu, proof burden-i də faydalı istiqamətə qaytarır. Şirkət clinical AI-sının care-i yaxşılaşdırdığını iddia etmək istəyirsə, uyğun sübut leaderboard deyil. Pasiyent üçün vacib outcome-larda belə trial-dır — idealda daha böyük, çünki burada dürüst dərs budur ki, mülayim faydaları görmək üçün böyük tədqiqatlar lazımdır.

Təmiz xülasə

Keniyada 16 primary care facility-də pragmatic, cluster-randomized trial clinical officers-un istifadə etdiyi electronic record-a əlavə edilən generative-AI decision-support alətini ("AI Consult") yoxladı. 9,691 pasiyent arasında 14 gün ərzində expert-adjudicated treatment-failure composite alətlə 2.2%, alətsiz 2.0% idi (adjusted odds ratio 0.77, 95% CI 0.55–1.08, P = 0.13) — əhəmiyyətli fərq yoxdur. Alət safety signal göstərmədi, bütün qiymətləndirilən sahələrdə clinical documentation-u yaxşılaşdırdı, prescribing-i dəyişmədi, patient satisfaction-a təsir etmədi və bir qədər aşağı antibiotic costs ilə əlaqəli idi. Müəlliflər bu limitlər daxilində təhlükəsiz göründüyünü, amma treatment failure azaltmadığını və hər hansı faydanın yəqin mülayim olduğunu nəticə çıxarır; müşahidə edilən ölçüdə effekti tapmaq çox daha böyük trial, təxminən 100,000 pasiyent miqyası tələb edə bilər. Nəticə clinical AI-nın patient outcomes-u yaxşılaşdırdığını və ya faydasız olduğunu göstərmir — bir urban private network-də təhlükə-

siz görünən və prosesi yaxşılaşdıran alətin iki həftədə pasiyentlərə nümayiş etdirilən fayda vermədiyini göstərir.

No-BS yoxlaması

Məqalə nəyi göstərir: Real-world randomized trial-da LLM-based decision-support-u primary care record-a əlavə etmək safety signal yaratmadı, clinical documentation quality-ni yaxşılaşdırdı, prescribing-i dəyişmədi və sərt 14-day patient composite treatment failure-i əhəmiyyətli dərəcədə azaltmadı.

Mümkündür, amma sübut olunmayıb: Alətin bu trial-ın görmək üçün çox kiçik olduğu real kiçik treatment-failure azalması yaratması; bütün xərclər sayıldıqda pul qənaət etməsi; daha yaxşı documentation-un zamanla daha yaxşı care-a çevrilməsi.

Nəyi göstərmir: Clinical AI-nın patient outcomes-u yaxşılaşdırdığını; nadir hadisələr üçün təhlükəsiz olduğunu; clinicians-i əvəz etdiyini və ya keçdiyini; nəticələrin rural və ya daha yüksək gəlirli settings-ə transfer olduğunu; cost saving-in müəyyənləşdiyini.

Əsas məhdudiyyətlər: Müşahidə ediləndən daha böyük effekt üçün powered idi (nadir outcomes çox böyük sample tələb edir); configuration error control arm-ı contaminasiya etdi və shared-network habits müqayisəni bulanıqlaşdırdı, hər ikisi fərqi sıfıra doğru çəkir; onsuz da yüksək standartlı bir private urban network; prespecified noninferiority və ya safety framework yox idi; qısa 14-day horizon.

Ümumi oxucu nə qədər əmin ola bilər? Bu trial-da alətin safety signal göstərmədiyinə və documentation-u yaxşılaşdırdığına yüksək. Burada 14-day patient outcomes-u nümayiş etdirilən şəkildə yaxşılaşdırmadığına yüksək. Patient-benefit intervention kimi “işləyir” və ya “uğursuzdur” hökmünə aşağı — sual həqiqətən açıqdır və daha böyük tədqiqat tələb edir. Düzgün mövqe: primary care-i dəyişdirən və ya məhv edən AI haqqında hökm yox, təhlükəsiz görünən və care process-i yaxşılaşdıran alət barədə diqqətli real-world nəticə.

Mənbələr

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