The Clean Paper

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

Öruggt og hjálplegt er ekki það sama og árangursríkt

Í slembuðri klínískri rannsókn í Kenía bætti gervigreindaraðstoð skjalfestingu og samræmi við leiðbeiningar í heilsugæslu, en aðalendapunktur fyrir sjúklinga batnaði ekki marktækt. Rannsóknin fann heldur ekkert nýtt öryggismerki, en það jafngildir ekki sönnun um öryggi. Þetta er niðurstaða um vinnuferli í tilteknu heilbrigðiskerfi, ekki almenn staðfesting á klínískri gervigreind.

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

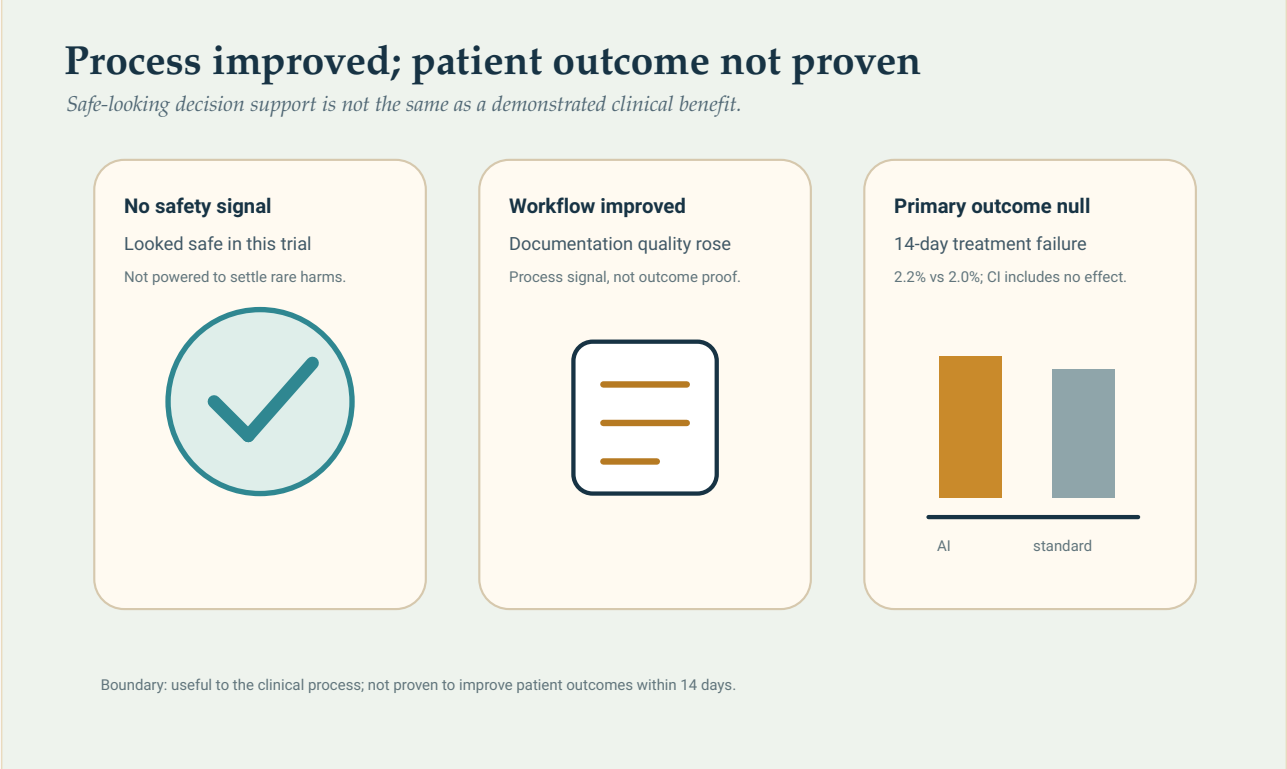
Öruggt og hjálplegt er ekki það sama og árangursríkt

Flestar fyrirsagnir um læknisfræðilega gervigreind byggja á röngum tegundum rannsókna. Líkan fær háa einkunn á prófspurningum eða vinnur lækna á vandlega völdum sjúkratilfellum og sagan skrifar sig sjálf: vélin er tilbúin á heilsugæsluna.

Þessi rannsókn gerði eitthvað erfiðara og sjaldgæfara. Hún setti ákvarðanastuðningskerfi með skapandi gervigreind inn í raunverulega heilsugæslu, á raunverulegum stöðvum, með raunverulegum sjúklingum, og spurði spurningarinnar sem skiptir máli: **varð sjúklingunum betra?**

Heiðarlega svarið er nei — ekki mælanlega innan 14 daga. Kerfið sýndi ekkert öryggismerki. Það bætti gæði klínískrar skráningar. Það kann jafnvel að hafa lækkað sum lyfjakostnað. En það minnkaði ekki með marktækum hætti meðferðarbrest, og höfundarnir segja varlega að ef ávinningur er til sé hann líklega hóflegur.

Það er ekki bilun rannsóknarinnar. Það er rannsóknin að virka. Svona líta ábyrg gögn um gervigreind í læknisfræði út þegar þau eru mæld á sjúklingum fremur en stigatöflum.



Kerfið sýndi ekkert öryggismerki og bætti klíníska skráningu, en fyrirfram skilgreinda 14 daga sjúklinganiðurstaðan batnaði ekki marktækt. Hjálp í ferli er ekki það sama og sannaður ávinningur fyrir sjúkling.

Hvað gerðu höfundarnir?

Teymið framkvæmdi hagnýta klasa-slembirannsókn á **16 heilsugæslustöðvum** í einkareknu heilbrigðiskerfi Penda Health í Nairobi- og Kiambu-sýslum í Kenía. Þjónustan er að stórum hluta veitt af **clinical officers**, millistigs heilbrigðisstarfsfólki með þriggja ára próf, oft án auðvelds aðgangs að sérfræðiráðgjöf.

Slembunin var á lækni/klínískum starfsmanni, ekki sjúklingi. **103 clinical officers** voru slembuð: 52 í íhlutunarhóp og 51 í samanburðarhóp. Báðir hópar notuðu sama skýjabyggða rafræna sjúkraskrárkerfi. Íhlutunarhópurinn fékk auk þess **AI Consult** (útgáfu 2.0), ákvarðanastuðning byggðan á **GPT-4o frá OpenAI** og samþættan skránni. Kerfið las það sem starfsmaðurinn skráði og gat bent á möguleg vandamál í greiningu eða meðferðaráætlun. Starfsfólkið hélt fullu ákvörðunarvaldi og gat samþykkt, breytt eða hunsað tillögur.

> **Hvaða líkan var notað og hvers vegna skiptir það máli?**
>
> Kerfið var **AI Consult 2.0**, sem notaði **GPT-4o frá OpenAI** (útgáfu maí 2025) í gegnum viðskipta-API OpenAI með enterprise-leyfi og með litlum tilviljanabreytileika (temperature 0,1). Það var inni í sérsniðnu EasyClinic rafrænu sjúkraskrárkerfi og kerfisfyrirmælin voru skrifuð til að samræmast kenískum meðferðarleiðbeiningum; höfundarnir birtu fyrirmælin í heild.
> Ástæðan fyrir þessum smáatriðum er að niðurstaðan snýr að **einu tilteknu kerfi** — einni útgáfu líkans, einum fyrirmælastreng, einni sjúkraskrá og einni stillingu — ekki „LLM í læknisfræði“ almennt. Höfundarnir segja það sama og kalla niðurstöðuna *tímabundið viðmið fremur en fasta getu*. Nýrra líkan, önnur fyrirmæli eða minna stafrænt heilbrigðiskerfi gæti gefið aðra niðurstöðu.
> Um sjálfstæði: OpenAI veitti síðar stuðning í fríðu, þar á meðal skýjakredit og tæknilega ráðgjöf um API-notkun. Höfundarnir segja að ákvörðunin um að nota OpenAI hafi verið tekin **áður** en það tilboð kom og að OpenAI hafi ekki tekið þátt í hönnun rannsóknarinnar, gagnasöfnun, greiningu eða ákvörðun um birtingu.

Milli 22. apríl og 16. júlí 2025 voru **9.691 sjúklingar** skráðir. **Aðalendapunkturinn var vísvitandi sjúklingamiðaður og strangur**: sérfræðingadæmd samsett mæling á *meðferðarbresti* innan 14 daga frá heimsókn — hópur klínískra sérfræðinga, blindur á rannsóknarhóp, mat hvort sjúklingur hefði haft slæma niðurstöðu eins og óleystan eða versnandi sjúkdóm. Rannsóknin var skráð fyrirfram (Pan-African Clinical Trials Registry 202502499779176).

Þessi hönnunarbreyta er kjarninn. Auðvelt er að sýna að gervigreind breyti því sem heilbrigðisstarfsmaður skrifar. Miklu erfiðara — og mikilvægara — er að sýna að hún breyti því sem gerist fyrir sjúklinginn.

Hvað fundu þeir?

Aðalendapunkturinn batnaði ekki. Meðferðarbrestur varð hjá 102 af 4.693 sjúklingum (2,2%) í AI-hópnum og 94 af 4.654 (2,0%) í samanburðarhópnum. Óleiðréttu prósenturnar voru örlítið hærri með AI, en eftir leiðréttingu hallaði punktmatið í átt að ávinningi: **leiðrétt odds ratio 0,77 (95% öryggisbil 0,55–1,08, P = 0,13)** — ekki tölfræðilega marktækt. Að hráa og leiðrétta matið halli í gagnstæðar áttir er ekki reiknivilla; leiðréttingin tekur tillit til munar milli klasa starfsfólks. Hvort heldur er inniheldur öryggisbilið greinilega „engin áhrif“, þannig að ekki er hægt að fullyrða um ávinning, og algildur munur var mjög lítill.

Fyrir einfaldari útskýringu á odds ratio, öryggisbilum og P-gildum sjá leiðbeiningu um klínískar niðurstöður.

Ekkert öryggismerki — innan marka rannsóknarinnar. Engir alvarlegir aukaatburðir voru taldir tengjast kerfinu og óháð mat fann ekkert öryggismerki. Höfundarnir eru skýrir um mörkin: rannsóknin var ekki nógu stór til að finna sjaldgæfa alvarlega skaða og hafði hvorki fyrirfram skilgreint noninferiority-markmið né formlegt öryggisramma. Hún getur því ekki *sannað* öryggi fyrir sjaldgæfa atburði.

Skráning varð betri. Í 2.000 heimsóknum sem blindaðir sérfræðingar yfirfóru var klínísk skráning betri í AI-hópnum í öllum metnum þáttum — greiningu, meðferðaráætlun og heildarfyllingu.

Lyfjaávísanir breyttust lítið. Enginn marktækur munur var á ávísunum, þar með talið réttri sýklalyfjanotkun (leiðrétt odds ratio 0,86, 95% CI 0,48–1,55). Kerfið breytti ekki tíðni sýklalyfjaávísana.

Sjúklingar tóku ekki eftir mun. Meðal 826 sjúklinga sem svöruðu ánægjukönnun var ánægja nánast eins milli hópa og heimsóknartími svipaður.

Kostnaður hallaði lítillega niður. Í leiðréttum greiningum var sýklalyfjatengdur kostnaður lægri í AI-hópnum — líklega vegna ódýrari valkosta fremur en færri ávísana — og sparnaður á sjúkling virtist meiri en rekstrarkostnaður kerfisins á sjúkling. Höfundarnir kalla þetta vísbendingu, ekki staðfesta niðurstöðu; full heildarkostnaðargreining var utan rannsóknarinnar.

Eigin einnar setningar samantekt höfundanna er hreinasta útgáfan: LLM-stuðningurinn sýndi ekkert öryggismerki innan þessara marka en minnkaði ekki meðferðarbrest innan 14 daga, og mögulegur ávinningur er líklega hóflegur.

Hvers vegna „enginn marktækur munur“ þýðir ekki „þetta virkar ekki“

Auðvelt er að oflesa null-niðurstöðu í báðar áttir. Tvö atriði stoppa einföldu söguna.

Í fyrsta lagi var **rannsóknin hönnuð til að finna stærri áhrif en þau sem sáust**. Alvarlegar slæmar niðurstöður í heilsugæslu eru sjaldgæfar — um 2% hér — þannig að það þarf gífurlega mörg tilvik til að greina lítinn raunverulegan ávinning frá hávaða. Eftir-á aflgreining höfundanna bendir til að áhrif

af þeirri stærð sem sást myndu krefjast **mun stærri rannsóknar, á stærðargráðunni yfir 100.000 sjúklingar**. Ómarktæk niðurstaða í þessari stærð útilokar því ekki lítinn raunverulegan ávinning; hún segir að rannsóknin gat ekki greint hann með vissu.

Í öðru lagi var **samanburðurinn að hluta mengaður**. Stillingarvilla gaf sumum starfsmönnum í samanburðarhópnum tímabundinn aðgang að AI Consult, og starfsmenn í sama neti tala saman og flytja vinnuvenjur milli hópa. Bæði atriði gera hópana líkari og draga raunverulegan mun að núlli. Þar að auki starfaði netið þegar við tiltölulega há gæðaviðmið, sem skilur minna svigrúm fyrir viðbótarverkfæri til að bæta árangur.

Ekkert af þessu bjargar fyrirsögn um „byltingu“. En það þýðir að rétta túlkunin er kvarðuð, ekki afskrift: á erfiðasta og heiðarlegasta endapunktinum hjálpaði kerfið ekki sjúklingum með sannanlegum hætti innan tveggja vikna — en það hjálpaði skráningu og virtist öruggt innan rannsóknarmarkanna.

Hvað sýnir þetta *ekki*?

- Það sýnir **ekki** að AI hafi bætt sjúklinganiðurstöður. Á aðal 14 daga endapunktinum var enginn marktækur ávinningur.
- Það sýnir **ekki** að AI sé gagnslaust. Punktmatið hallaði að ávinningi, skráning batnaði og lyfjakostnaður hallaði niður; null-niðurstaðan er samrýmanleg litlum raunverulegum ávinningi sem rannsóknin var of lítil til að staðfesta.
- Það **sannar ekki öryggi** gagnvart sjaldgæfum skaða. Ekkert öryggismerki fannst, en rannsóknin var ekki hönnuð til að votta öryggi fyrir sjaldgæfa alvarlega atburði.
- Það sýnir **ekki** að „AI vinni lækna“ eða geti komið í stað þeirra. Þetta er **ákvarðanastuðningur**; starfsmaðurinn hélt fullu valdi til að samþykkja eða hafna tillögum.
- Það **alhæfist ekki sjálfkrafa**. Rannsóknin fór fram í einu einkareknu borgarneti í Kenía. Dreifbýli, jaðarsvæði eða hátekjulönd gætu gefið aðrar niðurstöður í báðar áttir.
- Það **staðfestir ekki kostnaðarsparnað**. Kostnaðarmerkið er vísbending, ekki full efnahagsgreining.

Hversu sterk eru gögnin?

Fyrir meginfullyrðinguna — engin sönnuð lækkun á skammtíma sjúklingatjóni — eru gögnin sterk **að hönnun** og viðeigandi hófleg **að niðurstöðu**. Framsýn, fyrirfram skráð, klasa-slembirannsókn með blindu, sjúklingamiðuðu samsettu endamarki er nálægt bestu raunheimsgögnum sem hægt er að afla fyrir svona kerfi. Hún segir mun meira en stig á prófspurningum eða vignetterannsókn.

Fyrir aukaniðurstöður — betri skráningu, óbreyttar ávísanir, svipaða ánægju og lægri sýklalyfjakostnað — eru gögnin góð en aukaatriði. Þau eru styðjandi merki, ekki aðalfyrirsögn, og bera sömu takmarkanir vegna mengunar og eins heilbrigðisnets.

Fyrir öryggi eru gögnin hughreystandi en takmörkuð: ekkert merki fannst, en rannsóknin var ekki gerð til að finna sjaldgæfa skaða.

Gagnlegasta afstaðan er hvorki „þetta virkar“ né „þetta mistókst“. Hún er: vönduð raunheimsrannsókn fann að þetta AI-kerfi sýndi ekkert öryggismerki og bætti ferli umönnunar án þess að sýna sjúklingaávinning innan tveggja vikna — og að miklu stærri rannsókn þyrfti til að finna hugsanlegan hóflegan ávinning.

Hvers vegna skiptir þetta máli?

Umræðan um læknisfræðilega gervigreind skortir nákvæmlega svona gögn. Til eru þúsundir greina þar sem líkön leysa próf eða jafna sérfræðinga á snyrtilegum tilvikum. Mun færri stórar, hagnýtar, slembirannsóknir mæla hvort raunverulegum sjúklingum líði betur. Þetta er ein þeirra, og niðurstaðan lendir á óglæsilegri sannleikssetningu: að standast prófið er ekki það sama og að hjálpa sjúklingi.

Þetta bil er öll sagan. Verkfæri getur verið raunverulega gagnlegt fyrir starfsfólk — skýrari nótur, lægri lyfjakostnaður, annað auga — og samt ekki hreyft harðan sjúklingaendapunkt á tveimur vikum. Bæði geta verið satt samtímis, og þroskað heilbrigðiskerfi þarf að halda báðum staðreyndum í huga.

Rannsóknin færir líka sönnunarbyrðina á gagnlegan stað. Ef fyrirtæki vill halda fram að klínísk AI bæti umönnun, þá eru viðeigandi gögn ekki stigatafla. Það er rannsókn eins og þessi, með endapunktum sem skipta sjúklinga máli — og helst stærri rannsókn, því heiðarlegi lærdómurinn hér er að hóflegur ávinningur krefst stórra úrtaka til að sjást.

Í stuttu máli

Hagnýt klasa-slembirannsókn á 16 heilsugæslustöðvum í Kenía prófaði skapandi-AI ákvarðanastuðning, **AI Consult**, sem var bætt við rafræna sjúkraskrá clinical officers. Meðal 9.691 sjúklings var sérfræðingadæmt samsett meðferðarbrestshlutfall innan 14 daga 2,2% með kerfinu og 2,0% án þess (leiðrétt odds ratio 0,77, 95% CI 0,55–1,08, P = 0,13) — enginn marktækur munur. Kerfið sýndi ekkert öryggismerki, bætti klíníska skráningu í öllum metnum þáttum, breytti ekki ávísunum eða sjúklingaánægju og tengdist nokkru lægri sýklalyfjakostnaði. Höfundarnir ályktа að það hafi verið öruggt innan þessara marka en hafi ekki minnkað meðferðarbrest; hugsanlegur ávinningur sé líklega hóflegur og til að greina áhrif af þeirri stærð sem sást þyrfti rannsókn á stærðargráðu **100.000 sjúklinga**. Niðurstaðan sýnir hvorki að klínísk AI bæti sjúklingaárangur né að hún sé gagnslaus — hún sýnir að kerfi sem virtist öruggt og bætti ferli umönnunar hjálpaði ekki sjúklingum með sannanlegum hætti innan tveggja vikna í einu einkareknu borgarneti.

Raunhæf yfirferð

Það sem greinin sýnir: Í raunheims slembirannsókn sýndi LLM-byggður ákvarðanastuðningur ekkert öryggismerki, bætti klíníska skráningu, breytti ekki ávísunum og minnkaði ekki marktækt strangan 14 daga samsettan endapunkt meðferðarbrests.

Það sem er líklegt en ekki sannað: Að kerfið hafi lítinn raunverulegan ávinning sem rannsóknin var of lítil til að greina; að það spari peninga þegar allur kostnaður er talinn; eða að betri skráning leiði með tíma til betri umönnunar.

Það sem greinin sýnir ekki: Að klínísk AI bæti sjúklingaárangur; að hún sé örugg fyrir sjaldgæfa atburði; að hún komi í stað eða vinni heilbrigðisstarfsfólk; að niðurstöðurnar flytjist sjálfkrafa yfir í dreifbýli eða hátekjulönd; eða að kostnaðarsparnaður sé staðfestur.

Helstu takmarkanir: Rannsóknin var aflstillt fyrir stærri áhrif en sáust; sjaldgæfur endapunktur krefst mjög stórs úrtaks; stillingarvilla mengaði samanburðarhóp og sameiginlegt starfsmannanet getur dregið hópana saman; eitt einkarekið borgarnet með þegar góð gæði; enginn fyrirfram skilgreindur noninferiority- eða öryggisrammi; aðeins 14 daga eftirfylgd.

Hversu mikið traust ætti almennur lesandi að bera til niðurstöðunnar? Mikið traust á því að kerfið sýndi ekkert öryggismerki í þessari rannsókn og bætti skráningu. Mikið traust á því að það bætti **ekki sannanlega** 14 daga sjúklinganiðurstöður hér. Lítið traust á einföldum dómi að það „virki“ eða „virki ekki“ sem sjúklingaíhlutun — sú spurning er enn opin og þarf mun stærri rannsókn. Rétta afstaðan er vönduð raunheimsniðurstaða um verkfæri sem hjálpaði ferlinu og virtist öruggt, ekki dómur um að AI umbreyti — eða eyðileggi — heilsugæslu.

Heimildir

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