

The Clean Paper

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

Kliiniline AI paistis ohutu ja parandas dokumentatsiooni — kuid patsienditulemusi ei parandanud

Pragmaatiline klastripõhine randomiseeritud uuring 16 Kenya esmatasandi raviasutuses testis generatiivse AI otsustustoe tööriista, mis lisati clinical officer'ite kasutatavasse haiguslukku. 9691 patsiendi seas oli range, ekspertide hinnatud 14 päeva raviebaõnnestumise koondnäitaja tööriistaga 2,2% ja ilma 2,0% (korrigeeritud šansside suhe 0,77, 95% CI 0,55–1,08, P = 0,13) — statistiliselt olulist erinevust ei olnud. Ohutussignaali ei ilmnenud, dokumentatsioon paranes kõigis hinnatud valdkondades, ravimite määramine ja rahulolu jäid samaks ning antibiootikumikulud kaldusid allapoole. See ei ole läbikukkunud uuring, vaid haruldaselt aus uuring — mõõdetud patsientide, mitte võrdlustestide peal — ja selle väheglamuurne järeldus on, et ohutuna paistnud ja protsessi aidanud tööriist ei aidanud patsiente kahe nädala jooksul tõendatavalt ning võimaliku kasu tuvastamiseks oleks vaja palju suuremat uuringut.

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

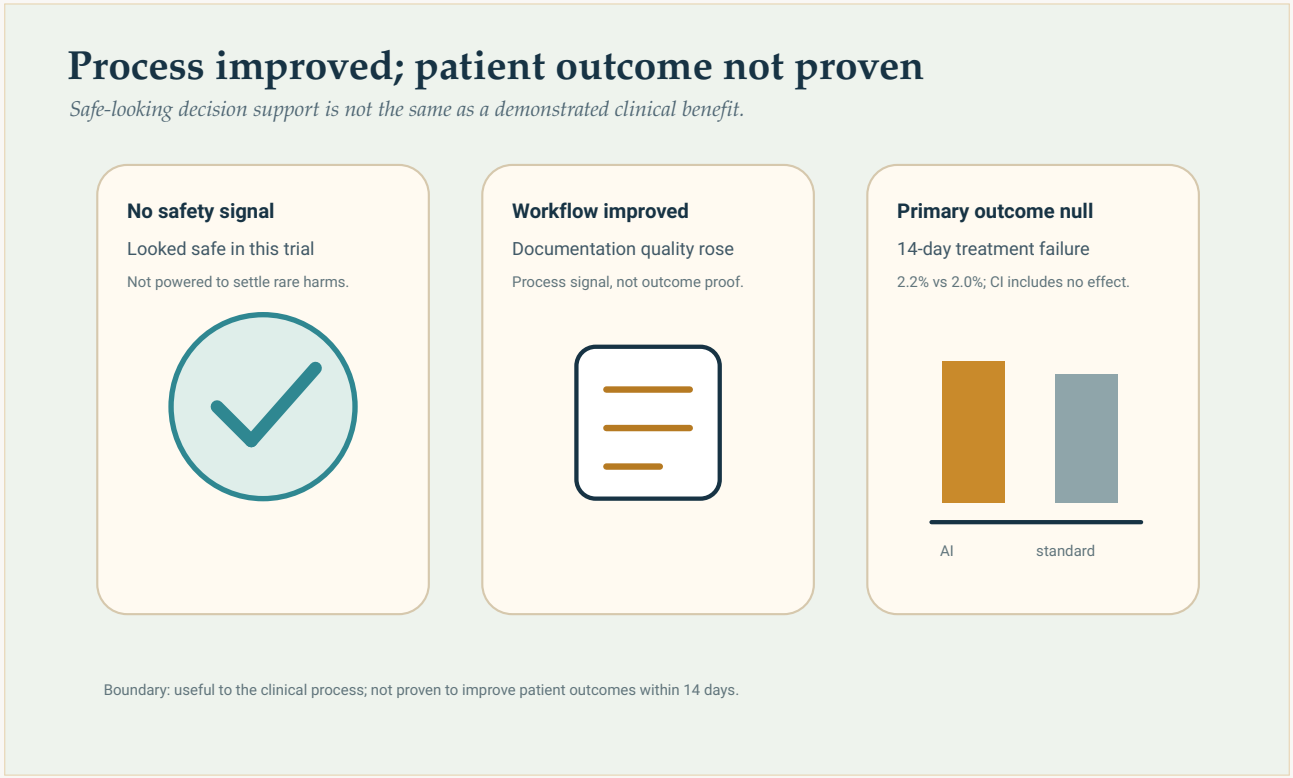
Ohutu ja abistav ei tähenda sama mis tõhus

Enamik meditsiinilise tehisintellekti pealkirju sünnib valest tüüpi uuringust. Mudel saab eksamiküsimustes häid tulemusi või edestab arste hoolikalt valitud haigusjuhtumites ning lugu kirjutab end ise: masin on kliinikuks valmis.

See uuring tegi midagi raskemat ja haruldasemat. Generatiivsel tehisintellektil põhinev otsustustoe tööriist viidi päris esmatasandi arstiabisse, päris raviasutustesse ja päris patsientide juurde ning küsiti küsimus, mis tegelikult loeb: kas patsientidel läks paremini?

Aus vastus on ei — vähemalt mitte mõõdetavalt, mitte 14 päeva jooksul. Tööriistaga ei ilmnenud ohutussignaali. See parandas kliinilise dokumentatsiooni kvaliteeti. Võimalik, et see vähendas isegi mõningaid ravimikulusid. Kuid raviebaõnnestumiste sagedus ei vähenenud statistiliselt oluliselt ning autorid ütlevad ettevaatlikult, et kui kasu üldse on, on see tõenäoliselt tagasihoidlik.

See ei ole uuringu läbikukkumine. See on uuring, mis teeb oma tööd. Nii näeb välja vastutustundlik tõendusmaterjal meditsiinilise tehisintellekti kohta, kui seda mõõdetakse patsientide, mitte võrdlustestide peal.



Tööriistaga ei ilmnenud ohutussignaali ja kliiniline dokumentatsioon paranes, kuid eelnevalt määratletud 14 päeva patsienditulemus ei paranenud statistiliselt oluliselt. Protsessi toetamine ei ole sama mis tõestatud kasu patsiendile.

The Clean Paper CC BY 4.0

Mida autorid tegid

Meeskond viis läbi pragmaatilise klastripõhise randomiseeritud uuringu **16 esmatasandi raviasutuses**, mida haldab eraõiguslik tervisevõrgustik Penda Health Kenya Nairobi ja Kiambu maakondades. Nendes asutustes osutavad ravi suuresti **clinical officers** — kolmeaastase diplomiga keskastme tervishoiutöötajad — kellel ei ole sageli lihtsat ligipääsu vanema kolleegi konsultatsioonile.

Randomiseerimise ühik oli klinitsist, mitte patsient. **103 clinical officer'it** randomiseeriti: 52 sekkumisrühma ja 51 kontrollrühma. Mõlemad rühmad kasutasid sama pilvepõhist elektroonilist terviselugu. Sekkumisrühmal oli lisaks **„AI Consult"** (versioon 2.0), **OpenAI GPT-4o** suurel keelemudelil põhinev otsustustoe tööriist, mis oli sellesse haiguslukku integreeritud. See luges klinitsisti dokumenteeritud teavet ning võis osutada võimalikele probleemidele diagnoosis või raviplaanis. Klinitsistidele jäi täielik otsustusõigus: nad võisid soovitused vastu võtta, neid muuta või eirata.

> **Milline mudel see oli ja miks üksikasjad loevad**
>
> Tööriist oli **AI Consult 2.0**, mis kasutas **OpenAI GPT-4o-d** (2025. aasta mai väljalase), millele pääseti ligi OpenAI kommerts-API kaudu ettevõttelitsentsi alusel ja mida käitati väikese juhuslikkusega (temperature 0.1). See asus kohandatud elektroonilises haigusloos (EasyClinicu EMR) ning seda suunasid süsteemiviibad, mis olid kirjutatud Kenya riiklike ravijuhistega kooskõlla; autorid avaldasid täieliku juhisviiba.
> Miks seda nii täpselt välja tuua? Sest tulemus puudutab *ühte konkreetset süsteemi* — üht mudeliversiooni, üht viipa, üht haiguslugu ja üht keskkonda — mitte „LLM-e meditsiinis" üldiselt. Autorid rõhutavad sama: nad nimetavad tulemust *ajaliseks võrdluspunktiks, mitte võimekuse püsivaks hinnanguks*. Uuem mudel, teine viip või vähem digitaliseeritud kliinik võiks tulemust muuta.
> Sõltumatuse kohta: OpenAI andis hiljem mitterahalist tuge (pilvandmetöötluse krediiti ja tehnilist juhendamist API kasutamisel), kuid autorid ütlevad, et otsus OpenAI-d kasutada tehti *enne* seda pakkumist ning OpenAI-l polnud rolli uuringu kavandamises, andmete kogumises, analüüsis ega avaldamisotsuses.

22. aprillist kuni 16. juulini 2025 kaasati **9691 patsienti**. **Esmane tulemusnäitaja oli teadlikult patsiendikeskne ja range**: ekspertide hinnatud *raviebaõnnestumise koondnäitaja* 14 päeva jooksul pärast visiiti — klinitsistide paneel hindas uuringurühma teadmata, kas patsiendil oli halb tulemus, näiteks haigus ei lahenenud või süvenes. Uuring registreeriti ette (Pan-African Clinical Trials Registry 202502499779176).

See disainivalik ongi asja mõte. Lihtne on näidata, et tehisintellekti tööriist muudab seda, mida klinitsist kirja paneb. Palju raskem ja palju tähenduslikum on näidata, et see muudab patsiendiga toimuvat.

Mida nad leidsid

Esmane tulemusnäitaja ei paranenud. Raviebaõnnestumine esines AI-rühmas 102 patsiendil 4693-st (2,2%) ja kontrollrühmas 94 patsiendil 4654-st (2,0%). Toorprotsendid olid tehisintellektiga pisut kõrgemad, kuid korrigeeritud punktihinnang kaldus kasu suunas: **korrigeeritud šansside suhe oli 0,77 (95% usaldusvahemik 0,55 kuni 1,08, P = 0,13)** — statistiliselt mitteoluline. Toor- ja korrigeeritud arvude vastassuunalisus ei ole arvutusviga; korrigeerimine arvestab erinevusi klinitsistide klastrite vahel. Mõlemal juhul sisaldab usaldusvahemik mugavalt „mõju puudumist", seega kasu väita ei saa, ning absoluutarvudes oli mõju väga väike.

Šansside suhte, usaldusvahemike ja P-väärtuste koos lugemise lihtsa selgituse leiad kliiniliste tulemuste lugemise juhendist.

Ohutussignaali ei ilmnenud — nende piiride sees. Ühtki tõsist kõrvaltoimet ei peetud tööriistaga seotuks ning sõltumatu ülevaatus ei leidnud ohutussignaali. Autorid on selle rahustava tulemuse piiride suhtes ausad: uuringul polnud piisavat võimsust haruldaste raskete kahjude leidmiseks ning selles polnud eelnevalt määratletud mittehalvemuse ega formaalset ohutusraamistikku, mistõttu see ei saa harvaesinevate sündmuste puhul ohutust *tõestada*.

Dokumentatsioon paranes. Pimedalt hinnatud ekspertide üle vaadatud 2000 visiidi hulgas koostasid AI Consulti kasutanud klinitsistid parema kliinilise dokumentatsiooni kõigis hinnatud valdkondades — diagnoosi kirjapanekus, raviplaanis ja üldises täielikkuses.

Ravimite määramine peaaegu ei muutunud. Ravimite määramises, sealhulgas antibiootikumide korrektses kasutuses, ei olnud statistiliselt olulist erinevust (korrigeeritud šansside suhe 0,86, 95% CI 0,48 kuni 1,55). Tööriist ei muutnud antibiootikumide väljakirjutamise määra.

Patsiendid erinevust ei märganud. Rahuloluküsitluse täitnud 826 patsiendi seas oli rahulolu rühmade vahel sisuliselt identne ning konsultatsioonide kestus oli sarnane.

Kulud viitasid väikesele langusele. Korrigeeritud analüüsis olid antibiootikumidega seotud kulud AI-rühmas väiksemad — tõenäoliselt odavamate valikute, mitte harvema väljakirjutamise tõttu — ning antibiootikumidelt patsiendi kohta säästetud summa näis ületavat tööriista kasutamise kulu patsiendi kohta. Autorid märgivad selle võimaliku, mitte lõpliku tulemusena: täielik omamiskulu analüüs jäi uuringust välja.

Autorite enda ühe lause kokkuvõte on kõige puhtam: LLM-i abi oli nende piiride sees ohutu, kuid ei vähendanud raviebaõnnestumist 14 päeva jooksul ning võimalik kasu on tõenäoliselt tagasihoidlik.

Miks „statistiliselt olulist erinevust ei olnud" ei tähenda „see ei tööta"

Nulltulemust esmase näitaja puhul on lihtne mõlemas suunas üle tõlgendada. Lihtsa loo takistavad kaks asja.

Esiteks **oli uuring kavandatud leidma suuremat mõju kui see, mida tegelikult nähti.** Tõsised halvad tulemused on esmatasandi arstiabis haruldased — siin umbes 2% — seega vajab väikese reaalse kasu eristamine mürast tohutuid valimeid. Autorite endi tagantjärele tehtud võimsusarvutused viitavad, et vaadeldud suurusega mõju leidmiseks oleks vaja **palju suuremat uuringut, suurusjärgus üle 100 000 patsiendi**. Mitteoluline tulemus sellise suurusega uuringus ei välista väikest tegelikku kasu; see tähendab, et uuring ei suutnud seda eristada.

Teiseks **oli võrdlus osaliselt hägustunud.** Seadistusviga andis mõnele kontrollrühma klinitsistile lühikeseks ajaks ligipääsu AI Consultile ning sama võrgustiku klinitsistid räägivad omavahel ja kannavad tööharjumusi üle rühmade piiri. Mõlemad mõjud muudavad rühmad sarnasemaks ja suruvad võimaliku tegeliku erinevuse nulli poole. Lisaks töötas vastuvõttev võrgustik juba suhteliselt kõrgete standardite järgi, mis jätab tööriistale vähem ruumi paranemist näidata.

Miski sellest ei päästa „läbimurde" pealkirja. Kuid see tähendab, et õige lugemine on kalibreeritud, mitte halvustav: kõige karmima ja ausama tulemusnäitaja järgi ei aidanud see tööriist patsiente kahe nädala jooksul tõendatavalt — samal ajal parandas see mõõdetavalt dokumenteerimist ja näis ohutu.

Mida see *ei* tõesta

- See **ei** näita, et AI parandas patsientide tulemusi. Esmase 14 päeva tulemusnäitaja järgi statistiliselt olulist kasu polnud.
- See **ei** näita, et AI oleks kasutu. Punktihinnang soosis seda, dokumentatsioon paranes kõikjal ja ravimikulud kaldusid allapoole; nulltulemus sobib kokku väikese reaalse kasuga, mille kinnitamiseks uuring oli liiga väike.
- See **ei** tõesta, et tööriist on haruldaste kahjude suhtes ohutu. Ohutussignaali ei leitud, kuid uuringul polnud piisavat võimsust ega sobivat disaini harvaesinevate raskete sündmuste ohutuse kinnitamiseks.
- See **ei** näita, et „AI võidab arste" või asendab klinitsiste. See on otsustus*tugi*; klinitsistile jäi täielik õigus soovitus vastu võtta või tagasi lükata.
- Seda **ei** saa automaatselt üldistada. Uuring toimus ühes Kenya erasektori linnavõrgustikus; maa-, linnalähedased ja kõrgema sissetulekuga keskkonnad võivad erineda mõlemas suunas.
- See **ei** tõesta kulude kokkuhoidu. Kulude signaal on vihjav, mitte valmis majandusanalüüs.

Kui tugevad on tõendid?

Keskse väite — lühiajalise patsiendikahju tõendatud vähenemist ei olnud — puhul on tõendusmaterjal *disaini poolest* tugev ja *järeldusena* sobivalt tagasihoidlik. Prospektiivne, eelregistreeritud, klastripõhine randomiseeritud uuring pimedalt hinnatud patsienditasandi koondnäitajaga on umbes parim reaaleluline tõendusmaterjal, mida sellise tööriista kohta koguda saab. See on palju informatiivsem kui võrdlustesti skoor või haigusjuhtumi vinjettide uuring.

Teiseste tulemuste — parem dokumentatsioon, muutumatu ravimite määramine, sarnane rahulolu,

väiksemad antibiootikumikulud — kohta on tõendid head, kuid neid tuleb lugeda teisestena: toetavad signaalid, mitte pealkiri, ning samade rühmadevahelise saastumise ja ühe võrgustiku piirangute suhtes haavatavad.

Ohutuse kohta on tõendid rahustavad, kuid piiratud: signaali ei leitud, kuid uuring polnud kavandatud haruldasi kahjusid leidma.

Kõige kasulikum hoiak ei ole ei „see töötab" ega „see kukkus läbi". See on: hoolikas reaaleluline uuring leidis, et AI-tööriistaga ei ilmnenud ohutussignaali ja raviprotsess paranes, kuid kahe nädala jooksul ei tõendatud patsienditulemuse kasu — ning sellise kasu leidmiseks oleks vaja palju suuremat uuringut.

Miks see oluline on

Meditsiinilise tehisintellekti arutelu kannatab just sellise tõendusmaterjali puuduse all. On tuhandeid töid, kus mudelid saavad eksamitel suurepäraseid tulemusi ja jõuavad korrastatud juhtumites klinitsistidega samale tasemele. Väga vähe on suuri pragmaatilisi randomiseeritud uuringuid, mis mõõdavad, kas päris patsientidel läheb paremini. See on üks neist ja jõuab väheglamuurse tõeni: testi läbimine ei ole sama mis patsiendi aitamine.

See vahe ongi kogu lugu. Tööriist võib olla klinitsistidele päriselt kasulik — selgemad märkmed, väiksemad ravimiarved, teine pilk — ja ikkagi mitte muuta kahe nädalaga ranget patsienditulemust. Mõlemad võivad korraga tõesed olla ning küps tervishoiusüsteem peab neid koos hoidma, mitte valima mugavamat.

See seab ka tõendamiskohustuse kasulikult paika. Kui ettevõte tahab väita, et tema kliiniline AI parandab ravi, pole asjakohane tõend edetabel. See on selline uuring, patsiendile oluliste tulemustega — ja ideaalis suurem, sest aus õppetund siin on, et tagasihoidliku kasu nägemiseks on vaja suuri uuringuid.

Lühidalt

Pragmaatiline klastripõhine randomiseeritud uuring 16 Kenya esmatasandi raviasutuses testis generatiivse AI otsustustoe tööriista („AI Consult"), mis lisati clinical officer'ite kasutatavasse elektroonilisse haiguslukku. 9691 patsiendi seas oli ekspertide hinnatud 14 päeva raviebaõnnestumise koondnäitaja tööriistaga 2,2% ja ilma 2,0% (korrigeeritud šansside suhe 0,77, 95% CI 0,55–1,08, P = 0,13) — statistiliselt olulist erinevust ei olnud. Tööriistaga ei ilmnenud ohutussignaali, kliiniline dokumentatsioon paranes kõigis hinnatud valdkondades, ravimite määramine ei muutunud, patsientide rahulolu jäi samaks ning antibiootikumikulud olid mõnevõrra väiksemad. Autorid järeldavad, et tööriist oli nende piiride sees ohutu, kuid ei vähendanud raviebaõnnestumist ning võimalik kasu on tõenäoliselt tagasihoidlik; vaadeldud suurusega mõju leidmiseks oleks vaja palju suuremat, suurusjärgus 100 000 patsiendiga uuringut. Tulemus ei näita, et kliiniline AI

parandab patsientide tulemusi, ega ka seda, et see on kasutu — see näitab, et ohutuna paistnud ja protsessi aidanud tööriist ei aidanud ühe erasektori linnavõrgustiku patsiente kahe nädala jooksul tõendatavalt.

Ilustamata kontroll

Mida töö näitab: Reaalelulises randomiseeritud uuringus ei tekitanud LLM-põhise otsustustoe lisamine esmatasandi arstiabi haiguslukku ohutussignaali, parandas kliinilise dokumentatsiooni kvaliteeti, ei muutnud ravimite määramist ega vähendanud statistiliselt oluliselt ranget 14 päeva raviebaõnnestumise koondnäitajat.

Mis on usutav, kuid tõestamata: Et tööriist vähendab raviebaõnnestumist vähesel määral, mida see uuring oli liiga väike tuvastama; et see säästab raha, kui arvestada kõiki kulusid; et parem dokumentatsioon muutub lõpuks paremaks raviks.

Mida see ei näita: Et kliiniline AI parandab patsientide tulemusi; et see on haruldaste sündmuste suhtes ohutu; et see asendab või edestab klinitsiste; et tulemused kanduvad üle maa- või kõrgema sissetulekuga keskkondadesse; et kulusääst on tõestatud.

Peamised piirangud: Uuringu võimsus oli kavandatud suurema mõju jaoks kui vaadeldud (haruldased tulemused vajavad väga suuri valimeid); seadistusviga saastas kontrollrühma ja sama võrgustiku ühised tööharjumused hägustasid võrdlust, mõlemad nihutavad tulemust erinevuse puudumise poole; üks juba kõrgete standarditega erasektori linnavõrgustik; puudus eelnevalt määratletud mittehalvemuse või ohutuse raamistik; lühike 14 päeva ajahorisont.

Kui palju võiks tavalugeja seda usaldada? Kõrge kindlus, et selles uuringus ei ilmnenud tööriistaga ohutussignaali ja dokumentatsioon paranes. Kõrge kindlus, et siin ei parandanud see tõendatavalt 14 päeva patsienditulemusi. Madal kindlus väidete suhtes, et see patsiendikasu sekkumisena kas „töötab" või „ei tööta" — see küsimus on päriselt lahtine ja vajab palju suuremat uuringut. Mõistlik hoiak: hoolikas reaaleluline tulemus tööriista kohta, mis ei tekitanud ohutussignaali ja parandas raviprotsessi, mitte otsus, et AI muudab — või rikub — esmatasandi arstiabi.

Allikad

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