The Clean Paper

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

Egy klinikai MI biztonságosnak látszott és javította a dokumentációt — de a betegkimeneteleket nem

Egy pragmatikus, klaszter-randomizált vizsgálat 16 kenyai alapellátási intézményben generatív MI-alapú döntéstámogató eszközt adott a clinical officerök nyilvántartásához. 9,691 beteg között a szigorú, 14 napos, szakértők által adjudikált összetett kezelési kudarc 2.2% volt az eszközzel és 2.0% nélküle (korrigált esélyhányados 0.77, 95% CI 0.55–1.08, P = 0.13) — nem szignifikáns különbség. Az eszköz nem mutatott biztonságossági jelet, minden területen javította a dokumentációt, nem változtatta meg a felírást vagy elégedettséget, és alacsonyabb antibiotikum-költség felé mutatott. Ez nem sikertelen, hanem ritka és őszinte vizsgálat: betegeken, nem benchmarkokon mérve azt találta, hogy egy biztonságosnak látszó, folyamatot javító eszköz két hét alatt nem adott kimutatható beteg-nyereséget, és egy esetleges kis előny felismeréséhez sokkal nagyobb vizsgálat kellene.

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

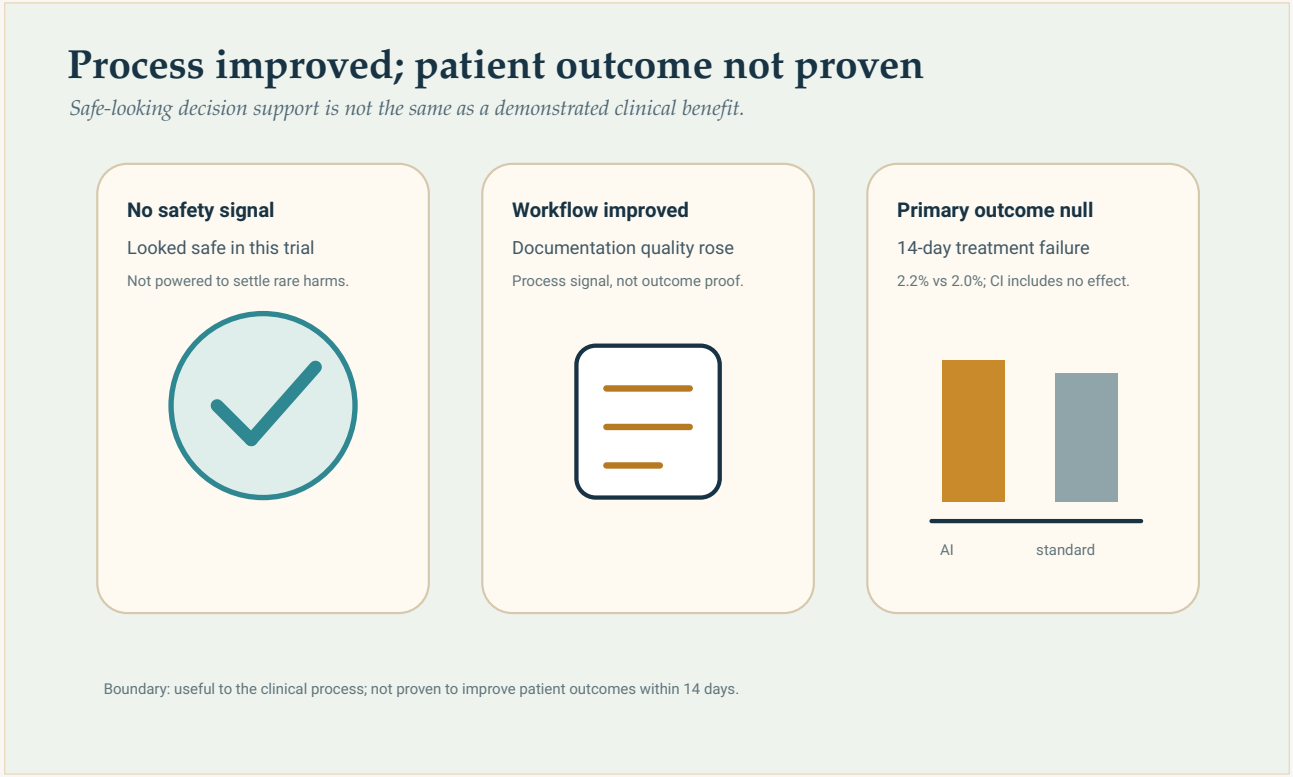
A biztonságos és hasznos nem ugyanaz, mint a hatásos

Az orvosi MI-ről szóló hírek többsége rossz típusú vizsgálatra épül. Egy modell jól teljesít vizsgakérdéseken vagy felülmúlja az orvosokat válogatott esetleírásokon, és a történet szinte megírja magát: a gép készen áll a klinikára.

Ez a vizsgálat nehezebb és ritkább dolgot tett. Generatív MI-alapú döntéstámogató eszközt vitt be valódi alapellátásba, valódi intézményekbe és valódi betegekhez, majd feltette azt a kérdést, amely ténylegesen számít: jobban jártak-e a betegek?

Az őszinte válasz: nem — legalábbis 14 napon belül mérhetően nem. Az eszköz nem mutatott biztonságossági figyelmeztető jelet. Javította a klinikai dokumentáció minőségét. Talán egyes gyógyszerköltségeket is csökkentett. A kezelési kudarcokat azonban nem csökkentette szignifikánsan, és a szerzők gondosan jelzik, hogy ha van is előnye, valószínűleg mérsékelt.

Ez nem a vizsgálat kudarca. Ez maga a vizsgálat működés közben. Így néz ki a felelős bizonyíték az orvosi MI-ről, amikor benchmarkok helyett betegeken mérik.



Az eszköz nem mutatott biztonságossági figyelmeztető jelet és javította a klinikai dokumentációt, de az előre meghatározott 14 napos betegkimenetel nem javult szignifikánsan. A folyamat támogatása nem ugyanaz, mint a bizonyított betegnyereség.

Mit csináltak a szerzők?

A csapat pragmatikus, klaszter-randomizált vizsgálatot végzett **16 alapellátási intézményben**, amelyeket a Penda Health magán-egészségügyi hálózat működtet Nairobi és Kiambu megyében, Kenyában. Az ellátást ezekben az intézményekben nagyrészt **clinical officerök** végzik — hároméves diplomával rendelkező középszintű klinikusok —, gyakran könnyen elérhető senior konzultáció nélkül.

A randomizálás egysége a klinikus volt, nem a beteg. **103 clinical officert** randomizáltak: 52-t az intervenciós, 51-et a kontrollkarba. Mindkét kar ugyanazt a felhőalapú elektronikus egészségügyi nyilvántartást használta. Az intervenciós kar ezen felül hozzáfért az **AI Consult** (2.0 verzió) döntéstámogató eszközhöz, amely **az OpenAI GPT-4o** nagy nyelvi modelljére épült, és a nyilvántartásba integrálták. Az eszköz elolvasta a klinikus által dokumentált információkat, és lehetséges problémákat jelezhetett a diagnózissal vagy kezelési tervvel kapcsolatban. A klinikusok teljes döntési autonómiát tartottak meg: elfogadhatták, módosíthatták vagy figyelmen kívül hagyhatták a javaslatokat.

> **Melyik modell volt, és miért számítanak a részletek?**
>
> Az eszköz **AI Consult 2.0** volt, **OpenAI GPT-4o** modellel (2025. májusi kiadás), az OpenAI kereskedelmi API-ján, vállalati licenc alatt, alacsony véletlenszerűségű beállítással (temperature 0.1). Egy egyedi elektronikus nyilvántartásba (EasyClinic EMR) épült, és a kenyai nemzeti kezelési irányelvekhez igazított rendszerpromptok vezérelték; a szerzők a teljes utasításpromptot közzétették.
> Miért fontos ezt ilyen pontosan leírni? Mert az eredmény *egy konkrét rendszerről* szól — egy modellverzióról, egy promptról, egy nyilvántartásról és egy környezetről —, nem általában „az LLM-ekről az orvoslásban”. A szerzők ugyanezt hangsúlyozzák: eredményüket *időbeli benchmarknak, nem rögzített képességbecslésnek* nevezik. Újabb modell, más prompt vagy kevésbé digitalizált klinika mind megváltoztathatná az eredményt.
> A függetlenségről: az OpenAI később természetbeni támogatást adott (felhőszámítási krediteket és technikai útmutatást az API használatához), de a szerzők szerint az OpenAI használatáról szóló döntés *e felajánlás előtt* megszületett, és az OpenAI-nak nem volt szerepe a vizsgálat tervezésében, adatgyűjtésében, elemzésében vagy a publikálási döntésben.

2025. április 22. és július 16. között **9,691 beteget** vontak be. Az **elsődleges kimenetel szándékosan betegközpontú és szigorú volt**: szakértők által adjudikált *összetett kezelési kudarc* a vizit utáni 14 napon belül — klinikusokból álló panel a vizsgálati kar ismerete nélkül ítélte meg, történt-e rossz kimenetel, például fennmaradó vagy romló betegség. A vizsgálatot előzetesen regisztrálták (Pan-African Clinical Trials Registry 202502499779176).

Ez a tervezési választás a lényeg. Könnyű megmutatni, hogy egy MI-eszköz megváltoztatja, mit ír le a klinikus. Sokkal nehezebb — és sokkal fontosabb — megmutatni, hogy megváltoztatja, mi történik a beteggel.

Mit találtak?

Az elsődleges kimenetel nem javult. Kezelési kudarc az MI-kar 4,693 betegéből 102-nél (2.2%), a kontrollkar 4,654 betegéből 94-nél (2.0%) történt. A nyers százalékok minimálisan magasabbak voltak MI mellett, de korrekció után a pontbecslés inkább előny felé mutatott: a **korrigált esélyhányados 0.77 volt (95%-os konfidenciaintervallum 0.55–1.08, P = 0.13)** — nem statisztikailag szignifikáns. A nyers és korrigált számok irányának megfordulása nem számítási hiba; a korrekció figyelembe veszi a klinikuscsoportok közötti különbségeket. A konfidenciaintervallum mindenesetre kényelmesen tartalmazza a „nincs hatás" lehetőségét, így előny nem állítható, abszolút értelemben pedig az eltérés kicsi volt.

Az esélyhányadosok, konfidenciaintervallumok és P-értékek együttes, hétköznapi értelmezéséhez lásd a klinikai eredmények olvasásáról szóló útmutatót.

Nem jelent meg biztonságossági jel — bizonyos határok között. Egyetlen súlyos nemkívánatos eseményt sem ítéltek az eszközzel összefüggőnek, és a független felülvizsgálat sem talált biztonságossági jelzést. A szerzők nyíltan jelzik ennek korlátját: a vizsgálatot nem ritka súlyos károk kimutatására méretezték, és nem volt előre meghatározott noninferiority vagy formális biztonságossági kerete, ezért ritka eseményekre nézve nem *bizonyíthatja* a biztonságosságot.

A dokumentáció javult. Kétezer, vak szakértők által felülvizsgált találkozásban az AI Consultot használó klinikusok minden értékelt területen jobb dokumentációt készítettek — diagnózis, kezelési terv és általános teljesség.

A gyógyszerfelírás alig mozdult. Nem volt szignifikáns különbség a felírásban, beleértve a helyes antibiotikum-használatot (korrigált esélyhányados 0.86, 95% CI 0.48–1.55). Az eszköz nem változtatta meg az antibiotikum-felírás arányát.

A betegek nem érzékeltek különbséget. A 826 elégedettségi kérdőívet kitöltő beteg között az elégedettség lényegében azonos volt a két karban, és a konzultációk hossza is hasonló maradt.

A költségek enyhén lefelé mutattak. Korrigált elemzésben az antibiotikumhoz kapcsolódó költségek alacsonyabbak voltak az MI-karban — valószínűleg olcsóbb gyógyszerválasztás, nem kevesebb recept miatt —, és a betegenkénti antibiotikum-megtakarítás látszólag meghaladta az eszköz betegenkénti futtatási költségét. A szerzők ezt jelzésnek, nem lezárt eredménynek tekintik: a teljes tulajdonlási költség gazdasági elemzése kívül esett a vizsgálaton.

A szerzők saját egymondatos összefoglalója a legtisztább: az LLM-asszisztencia e korlátok között biztonságosnak látszott, de 14 napon belül nem csökkentette a kezelési kudarcot, és minden esetleges előny valószínűleg szerény.

Miért nem ugyanaz a „nincs szignifikáns különbség”, mint az, hogy „nem működik”?

A null elsődleges kimenetelt könnyű mindkét irányba túlértelmezni. Két dolog akadályozza meg az egyszerű történetet.

Először: **a vizsgálatot nagyobb hatás kimutatására tervezték, mint amit megfigyeltek.** A súlyos rossz kimenetelek az alapellátásban ritkák — itt körülbelül 2% —, ezért egy kis valós előnyt a zajtól csak nagyon nagy mintában lehet elkülöníteni. A szerzők utólagos teljesítményszámításai szerint a megfigyelt nagyságú hatás kimutatásához **jóval nagyobb, nagyságrendileg több mint 100,000 beteget bevonó vizsgálat** kellene. Egy ekkora tanulmány nem szignifikáns eredménye nem zár ki kis valós előnyt; azt jelenti, hogy ezt a tanulmány nem tudta feloldani.

Másodszor: **az összehasonlítás részben elmosódott.** Egy konfigurációs hiba rövid ideig néhány kontrollkaros klinikusnak is hozzáférést adott az AI Consulthoz, a közös hálózatban dolgozó klinikusok pedig beszélnek egymással és szokásokat vihetnek át a csoporthatáron. Mindkettő hasonlóbbá teszi a két kart, és a valós különbséget nulla felé tolhatja. Emellett a szolgáltatóhálózat eleve viszonylag magas standardok szerint működött, így kevesebb tér maradt javulásra.

Egyik sem menti meg a „áttörés” címet. A helyes olvasat azonban kalibrált, nem lemondó: a legnehezebb és legőszintébb végponton ez az eszköz két héten belül nem mutatott kimutatható betegnyereséget — miközben mérhetően javította a dokumentációt és biztonságosnak látszott.

Mit *nem* bizonyít ez?

- **Nem** mutatja, hogy az MI javította a betegkimeneteleket. A fő 14 napos végponton nem volt szignifikáns előny.
- **Nem** mutatja, hogy az MI haszontalan. A pontbecslés előny felé mutatott, a dokumentáció minden területen javult, a gyógyszerköltségek pedig lefelé tendáltak; a null eredmény összeegyeztethető kis valós előnnyel, amelyhez ez a vizsgálat túl kicsi volt.
- **Nem** bizonyítja a ritka károkra vonatkozó biztonságosságot. Nem talált biztonságossági jelet, de nem ritka súlyos események igazolására tervezték.
- **Nem** mutatja, hogy „az MI jobb az orvosoknál” vagy helyettesíti őket. Ez döntés*támogatás*; a klinikus teljes jogkört tartott meg az elfogadásra vagy elutasításra.
- **Nem** általánosítható automatikusan. Egyetlen kenyai városi magánhálózatban zajlott; vidéki, periurbánus vagy magasabb jövedelmű környezet más eredményt adhat bármelyik irányban.
- **Nem** állapít meg biztos költségmegtakarítást. A költségjelzés érdekes, nem teljes gazdasági értékelés.

Mennyire erős a bizonyíték?

A központi állításra — nincs bizonyított csökkenés a rövid távú betegkárosodásban — a bizonyíték a *tervezés* szempontjából erős, a *következtetés* pedig helyesen szerény. Prospektív, előzetesen regisztrált, klaszter-randomizált vizsgálat vak, betegszintű összetett kimenetellel közel áll ahhoz a legjobb valós világban megszerezhető bizonyítékhoz, amely egy ilyen eszközről gyűjthető. Sokkal informatívabb benchmarkpontszámnál vagy esetleírásos vizsgálatnál.

A másodlagos eredményekre — jobb dokumentáció, változatlan felírás, hasonló elégedettség, alacsonyabb antibiotikum-költség — jó a bizonyíték, de másodlagosként kell olvasni: támogató jelek, nem a főcím, és ugyanazok a kontaminációs és egyetlen hálózatra vonatkozó korlátok érintik őket.

Biztonságosság szempontjából a bizonyíték megnyugtató, de behatárolt: nem találtak jelet, de a vizsgálat nem ritka károk felismerésére épült.

A leghasznosabb hozzáállás sem az, hogy „működik", sem az, hogy „megbukott". Hanem: egy gondos valós vizsgálat szerint ez az MI-eszköz nem mutatott biztonságossági figyelmeztető jelet és javította az ellátási folyamatot, de két hét alatt nem bizonyított betegkimeneteli előnyt — és egy esetleges ilyen előny felismeréséhez sokkal nagyobb vizsgálat kellene.

Miért fontos?

Az orvosi MI-ről szóló vitából éppen ez a fajta bizonyíték hiányzik. Tanulmányok ezrei mutatják, hogy modellek jól vizsgáznak és rendezett eseteken utolérik a klinikusokat. Nagy, pragmatikus, randomizált vizsgálatból, amely azt méri, valódi betegek jobban járnak-e, nagyon kevés van. Ez egy közülük, és a kevéssé csillogó igazságra jut: a teszt teljesítése nem ugyanaz, mint a beteg segítése.

Ez a rés az egész történet. Egy eszköz valóban hasznos lehet a klinikusoknak — tisztább dokumentáció, alacsonyabb gyógyszerszámla, második szem —, és közben két hét alatt mégsem mozdít el egy kemény betegkimenetelt. Mindkettő egyszerre lehet igaz, és egy érett egészségügyi rendszernek együtt kell tartania őket, nem a kényelmesebbet kiválasztani.

A bizonyítási terhet is hasznos irányba állítja vissza. Ha egy vállalat azt állítja, hogy klinikai MI-je javítja az ellátást, a releváns bizonyíték nem egy ranglista. Hanem egy ilyen vizsgálat, a betegeknek számító kimenetelekkel — és ideális esetben ennél nagyobb, mert az őszinte lecke itt az, hogy szerény előnyök meglátásához nagy vizsgálatok kellenek.

Röviden

Egy pragmatikus, klaszter-randomizált vizsgálat 16 kenyai alapellátási intézményben generatív MI-alapú döntéstámogató eszközt („AI Consult") adott a clinical officerök elektronikus nyilvántartásához. **9,691 beteg** között a 14 napon belüli, szakértők által adjudikált összetett kezelési kudarc 2.2% volt

az eszközzel, és 2.0% nélküle (korrigált esélyhányados 0.77, 95% CI 0.55–1.08, P = 0.13) — nem szignifikáns különbség. Az eszköz nem mutatott biztonságossági jelet, minden értékelt területen javította a klinikai dokumentációt, nem változtatta meg a gyógyszerfelírást, a beteg-elégedettséget sem, és valamivel alacsonyabb antibiotikum-költséggel társult. A szerzők szerint e határok között biztonságosnak látszott, de nem csökkentette a kezelési kudarcot; minden előny valószínűleg mérsékelt, és a megfigyelt nagyságú hatáshoz nagyságrendileg 100,000 beteget érintő vizsgálat kellene. Az eredmény nem mutatja, hogy a klinikai MI javítja a betegkimeneteleket, de azt sem, hogy haszontalan — azt mutatja, hogy egy biztonságosnak látszó és folyamatot javító eszköz két héten belül nem adott kimutatható betegnyereséget egyetlen városi magánhálózatban.

Tárgyilagos mérleg

Amit a tanulmány megmutat: Valós környezetben végzett randomizált vizsgálatban egy LLM-alapú döntéstámogató eszköz hozzáadása az alapellátási nyilvántartáshoz nem mutatott biztonságossági figyelmeztető jelet, javította a klinikai dokumentáció minőségét, nem változtatta a felírást, és nem csökkentette szignifikánsan a szigorú 14 napos összetett beteg-kudarcvégpontot.

Ami valószínű, de nincs bizonyítva: Hogy az eszköz kismértékben valóban csökkenti a kezelési kudarcot, csak a vizsgálat túl kicsi volt ennek kimutatására; hogy a teljes költségeket számolva pénzt takarít meg; hogy a jobb dokumentáció idővel jobb ellátássá alakul.

Amit nem mutat meg: Hogy a klinikai MI javítja a betegkimeneteleket; hogy ritka események szempontjából biztonságos; hogy helyettesíti vagy felülmúlja a klinikusokat; hogy az eredmények vidéki vagy magasabb jövedelmű környezetbe átvihetők; vagy hogy a költségmegtakarítás bizonyított.

Fő korlátok: A vizsgálatot a megfigyeltnél nagyobb hatásra méretezték — ritka kimenetekhez nagyon nagy minták kellenek; konfigurációs hiba kontaminálta a kontrollkart, a közös hálózatban terjedő szokások pedig szintén elmoshatják a különbséget, mindkettő nulla felé torzít; egyetlen, eleve magas standardú városi magánhálózat; nincs előre meghatározott noninferiority vagy formális biztonságossági keret; rövid, 14 napos horizont.

Mekkora bizalommal kezelje ezt egy általános olvasó? Nagy bizalommal azt, hogy ebben a vizsgálatban nem jelent meg biztonságossági jel, és a dokumentáció javult. Nagy bizalommal azt is, hogy itt nem volt kimutatható 14 napos betegkimeneteli javulás. Kis bizalommal minden olyan kategorikus állítást, hogy betegnyereség szempontjából „működik” vagy „nem működik” — ez valóban nyitott kérdés, amely sokkal nagyobb vizsgálatot igényel. A megfelelő hozzáállás: gondos valós eredmény egy olyan eszközről, amely nem mutatott biztonságossági jelet és javította az ellátási folyamatot, nem ítélet arról, hogy az MI átalakítja — vagy tönkreteszi — az alapellátást.

Források

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