The Clean Paper

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

Klinická AI vyzerala bezpečne a zlepšila dokumentáciu – ale nie výsledky pacientov

Pragmatická, klastrovo randomizovaná štúdia v 16 kenských zariadeniach primárnej starostlivosti testovala generatívny nástroj AI na podporu rozhodovania, vložený do elektronického zdravotného záznamu. U 9 691 pacientov bola odborne posúdená miera zlyhania liečby do 14 dní 2,2 % s nástrojom a 2,0 % bez neho (upravený pomer šancí 0,77; 95 % CI 0,55–1,08; P = 0,13) – bez významného rozdielu. Nástroj nevykázal bezpečnostný signál, zlepšil dokumentáciu, nezmenil predpisovanie ani spokojnosť pacientov a súvisel s nižšími nákladmi na antibiotiká. Štúdia merala výsledky pacientov, nie skóre v benchmarku, a priniesla triezvy záver: nástroj, ktorý sa javí ako bezpečný a pomáha procesu starostlivosti, počas dvoch týždňov nepreukázal prínos pre pacientov; zachytenie prípadného malého účinku by si vyžadovalo oveľa väčšiu štúdiu.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · 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

Bezpečný a užitočný neznamená efektívne

Väčšina titulkov o lekárskej umelej inteligencii vychádza z nesprávneho typu výskumu. Model dobre odpovedá na skúšobné otázky alebo porazí lekára na starostlivo vybraných kazuistikách a príbeh sa píše sám: stroj je pripravený do klinickej praxe.

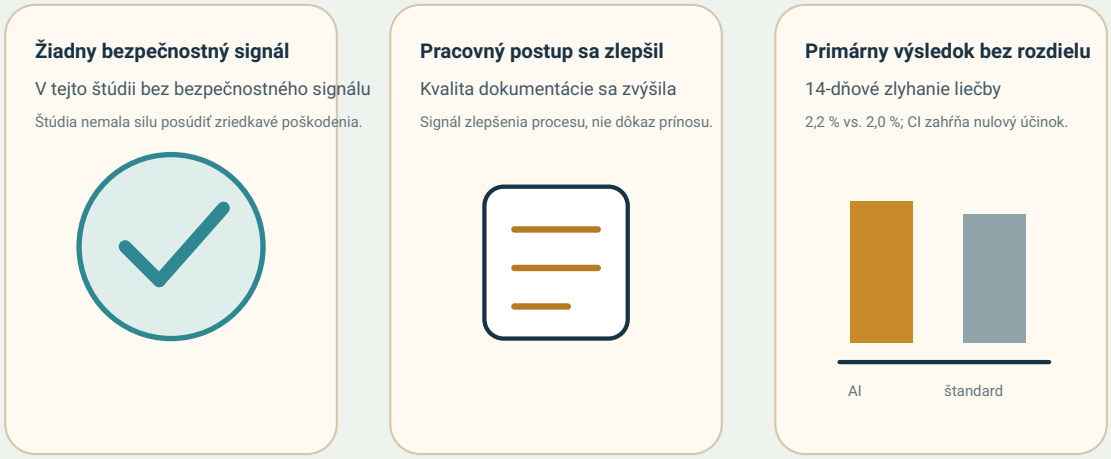
Táto štúdia urobila niečo ťažšie a vzácnejšie. Zaviedla generatívny nástroj AI na podporu rozhodovania do skutočnej primárnej starostlivosti, v skutočných zariadeniach a u skutočných pacientov, a položila otázku, na ktorej naozaj záleží: viedlo jeho použitie k lepším výsledkom pacientov?

Poctivá odpoveď znie nie – aspoň nie merateľne počas 14 dní. Nástroj nevykázal žiadny bezpečnostný signál. Zlepšil kvalitu klinickej dokumentácie a možno znížil niektoré náklady na lieky. Zlyhanie liečby však významne neobmedzil a autori upozorňujú, že prípadný prínos bude pravdepodobne mierny.

Nejde o zlyhanie štúdie. Naopak, takto má štúdia fungovať. Tak vyzerajú zodpovedné dôkazy o AI v medicíne, keď sa merajú výsledky pacientov namiesto výkonu v benchmarkoch.

Proces sa zlepšil; prínos pre pacientov nepreukázaný

Podpora rozhodovania bez bezpečnostného signálu nie je to isté ako preukázaný klinický prínos.



Hranica: užitočné pre klinický proces; zlepšenie výsledkov pacientov do 14 dní sa nepreukázalo.

Nástroj nevykazoval žiadny signál nebezpečenstva a zlepšil klinickú dokumentáciu, ale vopred definovaný 14-dňový výsledok pacienta sa významne nezlepšil. Podpora procesu nie je to isté ako preukázaný prínos pre pacienta.

Čo autori urobili

Tím vykonal pragmatickú, klastrovo randomizovanú štúdiu v **16 zariadeniach primárnej starostlivosti** súkromnej siete Penda Health v kenských okresoch Nairobi a Kiambu. Starostlivosť tu zaisťujú predovšetkým **clinical officers**, zdravotnícki pracovníci strednej úrovne s trojročným diplomom, často bez jednoduchého prístupu ku konzultácii so skúsenejším špecialistom.

Jednotkou randomizácie bol zdravotník, nie pacient. **103 clinical officers** bolo náhodne rozdelených: 52 do intervenčnej a 51 do kontrolnej skupiny. Obe skupiny používali ten istý cloudový elektronický zdravotný záznam. Intervenčná skupina mala navyše nástroj **AI Consult** vo verzii 2.0, založený na veľkom jazykovom modeli **GPT-4o od OpenAI** a vložený priamo do záznamu. Čítal údaje zapísané zdravotníkom a mohol upozorniť na možné problémy s diagnózou alebo liečebným plánom. Zdravotníci si zachovali plnú autonómiu: návrhy mohli prijať, upraviť alebo ignorovať.

> **Aký to bol model a prečo na detailoch záleží**
>
> Nástroj bol **AI Consult 2.0**, bežiaci na **GPT-4o od OpenAI** (vydanie z mája 2025), prístupný cez komerčné API OpenAI pod podnikovou licenciou a nastavený na nízku náhodnosť (teplota 0,1). Bol začlenený do špeciálneho elektronického záznamu (EMR EasyClinic) a riadil sa systémovými výzvami napísanými v súlade s národnými kenskými liečebnými smernicami; autori zverejnili celý návod na použitie.
> Prečo špecifikovať? Pretože výsledok je pre *jeden konkrétny systém* – jednu verziu modelu, výzvy, záznamu a prostredia – nie „LLM v medicíne“ všeobecne. Autori kladú dôraz na to isté: výsledok nazývajú *časový referenčný bod, nie trvalý odhad schopností*. Novší model, iná výzva alebo menej digitalizovaná klinika môže zmeniť výsledok.
> Pokiaľ ide o nezávislosť: OpenAI neskôr poskytlo vecnú podporu (kredity cloud computingu a technické pokyny k API), ale autori uvádzajú, že rozhodnutie použiť OpenAI bolo urobené *pred* touto ponukou a spoločnosť sa nijako nepodieľala na návrhu štúdie, zbere a analýze dát ani na rozhodnutí o zverejnení.

Od 22. apríla do 16. júla 2025 bolo zaradených **9 691 pacientov**. **Primárny výsledok bol zámerne zameraný na pacienta a prísne definovaný**: odborne posúdený *zložený ukazovateľ zlyhania liečby* do 14 dní od návštevy. Panel lekárov, ktorý nevedel, do ktorej skupiny pacient patril, hodnotil, či nastal nepriaznivý výsledok, napríklad pretrvávajúce alebo zhoršujúce sa ochorenie. Štúdia bola vopred registrovaná v Panafrickom registri klinických štúdií pod číslom 202502499779176.

Kľúčový je výber dizajnu. Je ľahké dokázať, že nástroj AI mení poznámky lekára. Je oveľa ťažšie – a oveľa dôležitejšie – ukázať, že to mení osud pacienta.

Čo objavili

Primárny výsledok sa nezlepšil. K zlyhaniu liečby došlo u 102 zo 4 693 pacientov (2,2 %) v skupine s AI a u 94 zo 4 654 (2,0 %) v kontrolnej skupine. Hrubé podiely boli pri AI nepatrne vyššie, ale po

štatistickej úprave bodový odhad smeroval k prínosu: **upravený pomer šancí predstavoval 0,77 (95 % interval spoľahlivosti 0,55–1,08; P = 0,13)**, teda bez štatistickej významnosti. Obrátenie smeru medzi hrubými a upravenými číslami nie je početnou chybou; úprava zohľadňuje rozdiely medzi skupinami zdravotníkov. V každom prípade interval spoľahlivosti pohodlne zahŕňa „žiadny účinok“, takže prínos tvrdiť nemožno a absolútny rozdiel bol nepatrný.

Jednoduché vysvetlenie spoločného čítania pomeru šancí, intervalu spoľahlivosti a hodnoty P ponúka sprievodca klinickými výsledkami.

Žiadny bezpečnostný signál – v rámci daných medzí. Žiadna závažná nežiaduca príhoda nebola posúdená ako súvisiaca s nástrojom a nezávislá kontrola žiadny bezpečnostný signál nenašla. Autori však uistenie poctivo ohraničujú: štúdia nemala štatistickú silu na odhalenie vzácnych závažných poškodení ani vopred stanovený rámec non-inferiority či formálneho hodnotenia bezpečnosti, takže bezpečnosť pri vzácnych udalostiach *dokázať* nemôže.

Dokumentácia sa zlepšila. V súbore 2 000 konzultácií hodnotených zaslepenými odborníkmi vytvárali zdravotníci používajúci AI Consult lepšiu dokumentáciu vo všetkých posudzovaných oblastiach – v zápise diagnózy, liečebnom pláne i celkovej úplnosti.

Predpisovanie sa takmer nezmenilo. Nebol žiadny významný rozdiel v predpisovaní, vrátane vhodného použitia antibiotík (upravený pomer šancí 0,86, 95 % CI 0,48–1,55). Tento nástroj nezmenil mieru predpisovania antibiotík.

Pacienti nezaznamenali rozdiel. Medzi 826 ľuďmi, ktorí dokončili prieskum spokojnosti, boli výsledky takmer totožné, rovnako ako doba konzultácie.

Náklady mierili mierne dole. V upravenej analýze boli náklady na antibiotiká v skupine s AI nižšie – zrejme vďaka voľbe lacnejších prípravkov, nie menšiemu počtu receptov – a úspora na pacienta sa zdala vyššia ako prevádzkové náklady nástroja. Autori to označujú za náznak, nie uzavretý záver: úplný výpočet celkových nákladov vlastníctva nebol súčasťou štúdie.

Zhrnutie autorov v jednej vete je najjasnejšie: LLM bola v rámci týchto limitov bezpečná, ale neznížila zlyhanie liečby do 14 dní a akýkoľvek prínos bude pravdepodobne mierny.

Prečo „žiadny významný rozdiel“ neznamená „nefunguje to“

Nulový hlavný výsledok je v oboch smeroch ľahko preinterpretovaný. Tomu bránia dve skutočnosti.

Po prvé, štúdia bola navrhnutá na zachytenie väčšieho účinku, než aký pozorovala. Závažné nepriaznivé výsledky sú v primárnej starostlivosti vzácne – tu okolo 2 % – takže na odlíšenie malého skutočného prínosu od šumu sú potrebné obrovské počty. Dodatočné výpočty štatistickej sily naznačujú, že zachytenie účinku tejto veľkosti by vyžadovalo **oveľa väčšiu štúdiu, rádovo cez 100 000 pacientov**. Nevýznamný výsledok v tejto vzorke nevylučuje malý skutočný prínos; znamená, že ho štúdia nedokázala rozlíšiť od náhody.

Po druhé, **porovnanie bolo čiastočne rozostrené**. Chyba konfigurácie krátko umožnila niektorým

zdravotníkom z kontrolnej skupiny prístup k AI Consult a pracovníci v spoločnej sieti spolu komunikujú a prenášajú návyky cez hranice skupín. Oba vplyvy skupiny zbližujú a prípadný skutočný rozdiel tlačí k nule. Sieť navyše už pred štúdiou udržovala pomerne vysoké štandardy, takže zostávalo menej priestoru na zlepšenie.

To nezachráni titulok o „prelome“. Znamená to však, že správny výklad má byť kalibrovaný, nie porazenecký: podľa najprísnejšieho a najpoctivejšieho cieľového ukazovateľa tento nástroj v priebehu dvoch týždňov preukázateľne nezlepšil výsledky pacientov, hoci merateľne zlepšil dokumentáciu a nevykázal bezpečnostný signál.

Čo to *nedokazuje*

- Neukazuje, že AI zlepšila výsledky pacientov. Primárny 14-dňový bod nevykazoval žiadny významný prínos.
- Neukazuje to, že AI je k ničomu. Bodový odhad bol v jej prospech, dokumentácia sa plošne zlepšovala a náklady na lieky klesali; nulový výsledok je v súlade s malým skutočným prínosom, ktorý štúdia nemohla potvrdiť.
- Nepreukazuje bezpečnosť nástroja pri zriedkavých poškodeniach. Žiadny bezpečnostný signál sa neobjavil, ale štúdia nebola navrhnutá ani dostatočne veľká na posúdenie vzácnych závažných udalostí.
- Neukazuje, že “AI poráža lekára” alebo ich nahrádza. *Podporuje* rozhodovanie; zdravotník si ponechal plnú právomoc návrh prijať alebo odmietnuť.
- Negeneralizuje automaticky. Skúška bola vykonaná v jednej súkromnej mestskej sieti v Keni; vidiecke, predmestské a vyššie príjmové prostredia sa môžu líšiť v oboch smeroch.
- Neurčuje úspory. Nákladový signál je sugestívny a nepredstavuje úplné ekonomické posúdenie.

Aké silné sú dôkazy?

Pri hlavnom závere – chýbajúcom preukaze zníženia krátkodobého poškodenia pacientov – sú dôkazy silné *dizajnom* a primerane skromné *záverom*. Prospektívna, vopred registrovaná, klastrovo randomizovaná štúdia so zaslepeným zloženým ukazovateľom na úrovni pacienta sa blíži najlepším dôkazom zo skutočného sveta, aké možno pre podobný nástroj získať. Hovorí oveľa viac ako skóre benchmarku alebo štúdie kazuistik.

Dôkazy pre sekundárne výsledky – lepšia dokumentácia, nezmenené predpisovanie, podobná spokojnosť, nižšie náklady na antibiotiká – sú dobré, ale treba ich chápať ako sekundárne: podporné signály, nie titulky, náchylné na kontamináciu z rovnakej skupiny a obmedzenie jednej siete.

Dôkazy o bezpečnosti sú upokojujúce, ale obmedzené: nebol nájdený žiadny signál, ale štúdia nebola vytvorená tak, aby odhalila vzácne poškodenia.

Najužitočnejší postoj nie je ani „funguje to“, ani „zlyhalo to“. Starostlivá štúdia v skutočnom svete

ukázala, že nástroj nevykázal bezpečnostný signál a zlepšil proces starostlivosti, ale v priebehu dvoch týždňov nepreukázal prínos pre výsledky pacientov – a zachytenie takého prínosu by vyžadovalo oveľa rozsiahlejšiu štúdiu.

Prečo je to dôležité

Debata o lekárskej AI má práve takých dôkazov zúfalo málo. Tisíce článkov ukazujú modely, ktoré zvládajú skúšky a dosahujú zhodu s lekármi pri prehľadných kazuistikách. Veľkých pragmatických randomizovaných štúdií merajúcich, či sa skutočným pacientom darí lepšie, existuje len veľmi málo. Toto je jedna z nich a prináša neokázalú pravdu: uspieť v teste nie je to isté ako pomôcť pacientovi.

V tejto medzere spočíva celý príbeh. Nástroj môže byť pre zdravotníkov skutočne užitočný – jasnejšie záznamy, nižšie účty za lieky, druhý pár očí – a napriek tomu počas dvoch týždňov nezmeniť závažné výsledky pacientov. Oboje môže platiť súčasne a vyspelý zdravotný systém musí udržať v zornom poli obe skutočnosti, nie si vybrať pohodlnejšie z nich.

Práca tiež vracia dôkazné bremeno na správne miesto. Ak chce firma tvrdiť, že jej klinická AI zlepšuje starostlivosť, relevantným dôkazom nie je rebríček. Je ním podobná štúdia zameraná na výsledky dôležité pre pacientov – a pokiaľ možno väčšie, pretože poctivé poučenie znie, že mierne prínosy vyžadujú rozsiahle štúdie.

Stručné zhrnutie

Pragmatická, klastrovo randomizovaná štúdia v 16 kenských zariadeniach primárnej starostlivosti testovala generatívny nástroj na podporu rozhodovania AI Consult, pridaný do elektronického zdravotného záznamu. Medzi 9 691 pacientmi bola odborne posúdená miera zlyhania liečby do 14 dní 2,2 % s nástrojom a 2,0 % bez neho (upravený pomer šancí 0,77; 95 % CI 0,55 – 1,08; P = 0,13) – bez významného rozdielu. Nástroj nevykázal bezpečnostný signál, zlepšil dokumentáciu vo všetkých hodnotených oblastiach, nezmenil predpisovanie ani spokojnosť pacientov a súvisel s mierne nižšími nákladmi na antibiotiká. Autori dospeli k záveru, že v medziach štúdie bol bezpečný, ale neznížil mieru zlyhania liečby a prípadný prínos bude pravdepodobne mierny; zachytenie účinku tejto veľkosti by vyžadovalo štúdiu rádovo so 100 000 pacientmi. Výsledok neukazuje ani to, že klinická AI zlepšuje výsledky pacientov, ani to, že je zbytočná. Ukazuje, že nástroj, ktorý sa javí ako bezpečný a pomáha procesu starostlivosti, počas dvoch týždňov nepriniesol preukázateľný prínos pre pacientov v jedinej súkromnej mestskej sieti.

Kontrola bez príkras

Čo práca ukazuje: V štúdii v skutočnej klinickej prevádzke pridanie nástroja na podporu rozhodovania založeného na LLM do záznamov primárnej starostlivosti nevykázalo bezpečnostný signál, zlepšilo kvalitu dokumentácie, nezmenilo predpisovanie a významne neznížilo prísne definovaný zložený ukazovateľ zlyhania liečby do 14 dní.

Čo je pravdepodobné, ale nepreukázané: Že nástroj môže mieru zlyhania liečby skutočne znížiť len nepatrne, príliš málo, aby to táto štúdia zachytila; že po započítaní všetkých nákladov šetrí peniaze; že sa lepšia dokumentácia časom premietne do lepšej starostlivosti.

Čo práca neukazuje: Že klinická AI zlepšuje výsledky pacientov; že nástroj je bezpečný aj z hľadiska zriedkavých udalostí; že nahrádza alebo prekonáva zdravotníkov; že výsledky možno preniesť do vidieckych alebo bohatších prostredí; že skutočne priniesol celkové úspory.

Hlavné obmedzenia: Štúdia mala štatistickú silu pre väčší účinok, než aký bol pozorovaný (zriedkavé výsledky vyžadujú rozsiahle štúdie); chyba konfigurácie čiastočne kontaminovala kontrolnú skupinu a zdieľané návyky v rámci siete rozostrili porovnanie – oba faktory tlačia výsledok k nulovému rozdielu; išlo o jedinú súkromnú mestskú sieť s už vysokými štandardmi; chýbal vopred stanovený rámec non-inferiority či formálneho hodnotenia bezpečnosti; horizont predstavoval len 14 dní.

Ako veľkú dôveru by mal mať čitateľ? Vysokú v záver, že nástroj v tejto vzorke nevykázal bezpečnostný signál a zlepšil dokumentáciu. Vysokú aj v tom, že tu nepreukázal zlepšenie 14-dňových výsledkov pacientov. Nízku pri kategorických tvrdeniach, že ako intervencia pre pacientov buď „funguje“, alebo „zlyhal“ – táto otázka zostáva otvorená a vyžaduje oveľa väčšiu štúdiu. Správny postoj: ide o starostlivý výsledok zo skutočnej prevádzky pri nástroji, ktorý nespôsobil bezpečnostný signál a zlepšil proces starostlivosti, nie o verdikt, že AI primárnu starostlivosť premieňa alebo ničí.

Zdroje

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