The Clean Paper

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

Klinická AI vypadala bezpečně a zlepšila dokumentaci – ale ne výsledky pacientů

Pragmatická, klastrově randomizovaná studie v 16 keňských zařízeních primární péče testovala generativní nástroj AI pro podporu rozhodování, vložený do elektronického zdravotního záznamu. U 9 691 pacientů činila odborně posouzená míra selhání léčby do 14 dnů 2,2 % s nástrojem a 2,0 % bez něj (upravený poměr šancí 0,77; 95% CI 0,55–1,08; P = 0,13) – bez významného rozdílu. Nástroj nevykázal bezpečnostní signál, zlepšil dokumentaci, nezměnil předepisování ani spokojenost pacientů a souvisel s nižšími náklady na antibiotika. Studie měřila výsledky pacientů, nikoli skóre v benchmarku, a přinesla střízlivý závěr: nástroj, který se jeví jako bezpečný a pomáhá procesu péče, během dvou týdnů neprokázal přínos pro pacienty; zachycení případného malého účinku by vyžadovalo mnohem větší studii.

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 užitečný neznamená efektivní

Většina titulků o lékařské umělé inteligenci vychází z nesprávného typu výzkumu. Model dobře odpovídá na zkušební otázky nebo porazí lékaře na pečlivě vybraných kazuistikách a příběh se píše sám: stroj je připraven do klinické praxe.

Tato studie udělala něco obtížnějšího a vzácnějšího. Zavedla generativní nástroj AI pro podporu rozhodování do skutečné primární péče, ve skutečných zařízeních a u skutečných pacientů, a položila otázku, na níž opravdu záleží: vedlo jeho použití k lepším výsledkům pacientů?

Poctivá odpověď zní ne – alespoň ne měřitelně během 14 dnů. Nástroj nevykázal žádný bezpečnostní signál. Zlepšil kvalitu klinické dokumentace a možná snížil některé náklady na léky. Selhání léčby však významně neomezil a autoři upozorňují, že případný přínos bude pravděpodobně mírný.

Nejde o selhání studie. Naopak, takto má studie fungovat. Tak vypadají odpovědné důkazy o AI v medicíně, když se měří výsledky pacientů namísto výkonu v benchmarcích.

Nástroj nevykazoval žádný signál nebezpečí a zlepšil klinickou dokumentaci, ale předem definovaný 14denní výsledek pacienta se významně nezlepšil. Podpora procesu není totéž jako prokázaný přínos pro pacienta.

Co autoři udělali

Tým provedl pragmatickou, klastrově randomizovanou studii v **16 zařízeních primární péče** soukromé sítě Penda Health v keňských okresech Nairobi a Kiambu. Péči zde zajišťují především **clinical officers**, zdravotničtí pracovníci střední úrovně s tříletým diplomem, často bez snadného přístupu ke konzultaci se zkušenějším specialistou.

Jednotkou randomizace byl zdravotník, nikoli pacient. **103 clinical officers** bylo náhodně rozděleno: 52 do intervenční a 51 do kontrolní skupiny. Obě skupiny používaly tentýž cloudový elektronický zdravotní záznam. Intervenční skupina měla navíc nástroj **AI Consult** ve verzi 2.0, založený na velkém jazykovém modelu **GPT-4o od OpenAI** a vložený přímo do záznamu. Četl údaje zapsané zdravotníkem a mohl upozornit na možné problémy s diagnózou nebo léčebným plánem. Zdravotníci si zachovali plnou autonomii: návrhy mohli přijmout, upravit nebo ignorovat.

> **Jaký to byl model a proč na detailech záleží**
>
> Nástroj byl **AI Consult 2.0**, běžící na **GPT-4o od OpenAI** (vydání z května 2025), přístupný přes komerční API OpenAI pod podnikovou licencí a nastavený na nízkou náhodnost (teplota 0,1). Byl začleněn do speciálního elektronického záznamu (EMR EasyClinic) a řídil se systémovými výzvami napsanými v souladu s národními keňskými léčebnými směrnicemi; autoři zveřejnili celý návod k použití.
> Proč specifikovat? Protože výsledek je pro *jeden konkrétní systém* – jednu verzi modelu, výzvy, záznamu a prostředí – nikoli „LLM v medicíně“ obecně. Autoři kladou důraz na totéž: výsledek nazývají *časový referenční bod, nikoli trvalý odhad schopností*. Novější model, jiná výzva nebo méně digitalizovaná klinika může změnit výsledek.
> Pokud jde o nezávislost: OpenAI později poskytlo věcnou podporu (kredity cloud computingu a technické pokyny k API), ale autoři uvádějí, že rozhodnutí použít OpenAI bylo učiněno *před* touto nabídkou a společnost se nijak nepodílela na návrhu zkušebních verzí, sběru a analýze dat ani na rozhodnutí o zveřejnění.

Od 22. dubna do 16. července 2025 bylo zařazeno **9 691 pacientů**. **Primární výsledek byl záměrně zaměřený na pacienta a přísně definovaný**: odborně posouzený *složený ukazatel selhání léčby* do 14 dnů od návštěvy. Panel lékařů, který nevěděl, do které skupiny pacient patřil, hodnotil, zda nastal nepříznivý výsledek, například přetrvávající nebo zhoršující se onemocnění. Studie byla předem registrována v Panafrickém registru klinických studií pod číslem 202502499779176.

Základem je výběr designu. Je snadné prokázat, že nástroj AI mění poznámky lékaře. Je mnohem obtížnější – a mnohem důležitější – ukázat, že to mění osud pacienta.

Co objevili

Primární výsledek se nezlepšil. K selhání léčby došlo u 102 ze 4 693 pacientů (2,2 %) ve skupině s AI a u 94 ze 4 654 (2,0 %) v kontrolní skupině. Hrubé podíly byly u AI nepatrně vyšší, ale po stati-

stické úpravě bodový odhad směřoval k přínosu: **upravený poměr šancí činil 0,77 (95% interval spolehlivosti 0,55–1,08; P = 0,13)**, tedy bez statistické významnosti. Obrácení směru mezi hrubými a upravenými čísly není početní chybou; úprava zohledňuje rozdíly mezi skupinami zdravotníků. V každém případě interval spolehlivosti pohodlně zahrnuje „žádný účinek“, takže přínos tvrdit nelze a absolutní rozdíl byl nepatrný.

Jednoduché vysvětlení společného čtení poměru šancí, intervalu spolehlivosti a hodnoty P nabízí průvodce klinickými výsledky.

Žádný bezpečnostní signál – v rámci daných mezí. Žádná závažná nežádoucí příhoda nebyla posouzena jako související s nástrojem a nezávislá kontrola žádný bezpečnostní signál nenašla. Autoři však ujištění poctivě ohraničují: studie neměla statistickou sílu k odhalení vzácných závažných poškození ani předem stanovený rámec non-inferiority či formálního hodnocení bezpečnosti, takže bezpečnost u vzácných událostí *prokázat* nemůže.

Dokumentace se zlepšila. V souboru 2 000 konzultací hodnocených zaslepenými odborníky vytvářeli zdravotníci používající AI Consult lepší dokumentaci ve všech posuzovaných oblastech – v zápisu diagnózy, léčebném plánu i celkové úplnosti.

Předepisování se téměř nezměnilo. Nebyl žádný významný rozdíl v předepisování, včetně vhodného použití antibiotik (upravený poměr šancí 0,86, 95% CI 0,48–1,55). Tento nástroj nezměnil míru předepisování antibiotik.

Pacienti nezaznamenali rozdíl. Mezi 826 lidmi, kteří dokončili průzkum spokojenosti, byly výsledky téměř totožné, stejně jako doba konzultace.

Náklady mířily mírně dolů. V upravené analýze byly náklady na antibiotika ve skupině s AI nižší – patrně díky volbě levnějších přípravků, nikoli menšímu počtu receptů – a úspora na pacienta se zdála vyšší než provozní náklady nástroje. Autoři to označují za náznak, nikoli uzavřený závěr: úplný výpočet celkových nákladů vlastnictví nebyl součástí studie.

Shrnutí autorů v jedné větě je nejjasnější: LLM byla v rámci těchto limitů bezpečná, ale nesnížila selhání léčby do 14 dnů a jakýkoli přínos bude pravděpodobně mírný.

Proč „žádný významný rozdíl“ neznamená „nefunguje to“

Nulový hlavní výsledek je v obou směrech snadno přeinterpretován. Tomu brání dvě skutečnosti.

Za prvé, **studie byla navržena k zachycení většího účinku, než jaký pozorovala**. Závažné nepříznivé výsledky jsou v primární péči vzácné – zde kolem 2 % –, takže k odlišení malého skutečného přínosu od šumu jsou potřeba obrovské počty. Dodatečné výpočty statistické síly naznačují, že zachycení účinku této velikosti by vyžadovalo **mnohem větší studii, řádově přes 100 000 pacientů**. Nevýznamný výsledek v tomto vzorku nevylučuje malý skutečný přínos; znamená, že jej studie nedokázala rozlišit od náhody.

Za druhé, **srovnání bylo částečně rozostřené**. Chyba konfigurace krátce umožnila některým zdravot-

níkům z kontrolní skupiny přístup k AI Consult a pracovníci ve společné síti spolu komunikují a přenášejí návyky přes hranice skupin. Oba vlivy skupiny sbližují a případný skutečný rozdíl tlačí k nule. Síť navíc už před studií udržovala poměrně vysoké standardy, takže zbývalo méně prostoru pro zlepšení.

To nezachrání titulek o „průlomu“. Znamená to však, že správný výklad má být kalibrovaný, nikoli poraženecký: podle nejpřísnějšího a nejpoctivějšího cílového ukazatele tento nástroj během dvou týdnů prokazatelně nezlepšil výsledky pacientů, ačkoli měřitelně zlepšil dokumentaci a nevykázal bezpečnostní signál.

Co to *nedokazuje*

- Neukazuje, že AI zlepšila výsledky pacientů. Primární 14denní bod nevykazoval žádný významný přínos.
- Neukazuje to, že AI je k ničemu. Bodový odhad byl v její prospěch, dokumentace se plošně zlepšovala a náklady na léky klesaly; nulový výsledek je v souladu s malým skutečným přínosem, který studie nemohla potvrdit.
- Neprokazuje bezpečnost nástroje u vzácných poškození. Žádný bezpečnostní signál se neobjevil, ale studie nebyla navržena ani dostatečně velká k posouzení vzácných závažných událostí.
- Neukazuje, že “AI poráží lékaře” nebo je nahrazuje. *Podporuje* rozhodnutí; lékař si ponechal plnou pravomoc návrh přijmout nebo odmítnout.
- Negeneralizuje automaticky. Zkouška byla provedena v jedné soukromé městské síti v Keni; venkovská, předměstská a vyšší příjmová prostředí se mohou lišit v obou směrech.
- Neurčuje úspory. Nákladový signál je sugestivní a nepředstavuje úplné ekonomické posouzení.

Jak silné jsou důkazy?

U hlavního závěru – chybějícího průkazu snížení krátkodobého poškození pacientů – jsou důkazy silné *designem* a náležitě skromné *závěrem*. Prospektivní, předem registrovaná, klastrově randomizovaná studie se zaslepeným složeným ukazatelem na úrovni pacienta se blíží nejlepším důkazům ze skutečného světa, jaké lze pro podobný nástroj získat. Říká mnohem více než skóre benchmarku nebo studie kazuistik.

Důkazy pro sekundární výsledky – lepší dokumentace, nezměněné předepisování, podobná spokojenost, nižší náklady na antibiotika – jsou dobré, ale je třeba je chápat jako sekundární: podpůrné signály, nikoli titulky, náchylné ke kontaminaci ze stejné skupiny a omezení jedné sítě.

Důkazy o bezpečnosti jsou uklidňující, ale omezené: nebyl nalezen žádný signál, ale studie nebyla vytvořena tak, aby odhalila vzácná poškození.

Nejužitečnější postoj není ani „funguje to“, ani „selhalo to“. Pečlivá studie ve skutečném světě ukázala, že nástroj nevykázal bezpečnostní signál a zlepšil proces péče, ale během dvou týdnů neprokázal

přínos pro výsledky pacientů – a zachycení takového přínosu by vyžadovalo mnohem rozsáhlejší studii.

Proč je to důležité

Debata o lékařské AI má právě takových důkazů zoufale málo. Tisíce článků ukazují modely, které zvládají zkoušky a dosahují shody s lékaři u přehledných kazuistik. Velkých pragmatických randomizovaných studií měřících, zda se skutečným pacientům daří lépe, existuje jen velmi málo. Toto je jedna z nich a přináší neokázalou pravdu: uspět v testu není totéž jako pomoci pacientovi.

V této mezeře spočívá celý příběh. Nástroj může být pro zdravotníky skutečně užitečný – jasnější záznamy, nižší účty za léky, druhý pár očí – a přesto během dvou týdnů nezměnit závažné výsledky pacientů. Obojí může platit současně a vyspělý zdravotní systém musí udržet v zorném poli obě skutečnosti, nikoli si vybrat pohodlnější z nich.

Práce také vrací důkazní břemeno na správné místo. Chce-li firma tvrdit, že její klinická AI zlepšuje péči, relevantním důkazem není žebříček. Je jím podobná studie zaměřená na výsledky důležité pro pacienty – a pokud možno větší, protože poctivé poučení zní, že mírné přínosy vyžadují rozsáhlé studie.

Stručné shrnutí

Pragmatická, klastrově randomizovaná studie v 16 keňských zařízeních primární péče testovala generativní nástroj pro podporu rozhodování AI Consult, přidaný do elektronického zdravotního záznamu. Mezi 9 691 pacienty činila odborně posouzená míra selhání léčby do 14 dnů 2,2 % s nástrojem a 2,0 % bez něj (upravený poměr šancí 0,77; 95% CI 0,55–1,08; P = 0,13) – bez významného rozdílu. Nástroj nevykázal bezpečnostní signál, zlepšil dokumentaci ve všech hodnocených oblastech, nezměnil předepisování ani spokojenost pacientů a souvisel s mírně nižšími náklady na antibiotika. Autoři dospěli k závěru, že v mezích studie byl bezpečný, ale nesnížil míru selhání léčby a případný přínos bude pravděpodobně mírný; zachycení účinku této velikosti by vyžadovalo studii řádově se 100 000 pacienty. Výsledek neukazuje ani to, že klinická AI zlepšuje výsledky pacientů, ani to, že je zbytečná. Ukazuje, že nástroj, který se jeví jako bezpečný a pomáhá procesu péče, během dvou týdnů nepřinesl prokazatelný přínos pro pacienty v jediné soukromé městské síti.

Kontrola bez příkras

Co práce ukazuje: Ve studii ve skutečném klinickém provozu přidání nástroje pro podporu rozhodování založeného na LLM do záznamů primární péče nevykázalo bezpečnostní signál, zlepšilo kvalitu dokumentace, nezměnilo předepisování a významně nesnížilo přísně definovaný

složený ukazatel selhání léčby do 14 dnů.

Co je pravděpodobné, ale neprokázané: Že nástroj může míru selhání léčby skutečně snížit jen nepatrně, příliš málo, aby to tato studie zachytila; že po započtení všech nákladů šetří peníze; že se lepší dokumentace časem promítne do lepší péče.

Co práce neukazuje: Že klinická AI zlepšuje výsledky pacientů; že je nástroj bezpečný i z hlediska vzácných událostí; že nahrazuje nebo překonává zdravotníky; že lze výsledky přenést do venkovských či bohatších prostředí; že skutečně přinesl celkové úspory.

Hlavní omezení: Studie měla statistickou sílu pro větší účinek, než jaký byl pozorován (vzácné výsledky vyžadují rozsáhlé studie); chyba konfigurace částečně kontaminovala kontrolní skupinu a sdílené návyky v rámci sítě rozostřily srovnání – oba faktory tlačí výsledek k nulovému rozdílu; šlo o jedinou soukromou městskou síť s již vysokými standardy; chyběl předem stanovený rámec non-inferiority či formálního hodnocení bezpečnosti; horizont činil jen 14 dnů.

Jak velkou důvěru by měl mít čtenář? Vysokou v závěr, že nástroj v tomto vzorku nevykázal bezpečnostní signál a zlepšil dokumentaci. Vysokou také v závěr, že zde neprokázal zlepšení 14denních výsledků pacientů. Nízkou v kategorická tvrzení, že jako intervence pro pacienty buď „funguje“, nebo „selhal“ – tato otázka zůstává otevřená a vyžaduje mnohem větší studii. Správný postoj: jde o pečlivý výsledek ze skutečného provozu u nástroje, který nezpůsobil bezpečnostní signál a zlepšil proces péče, nikoli o verdikt, že AI primární péči proměňuje nebo 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