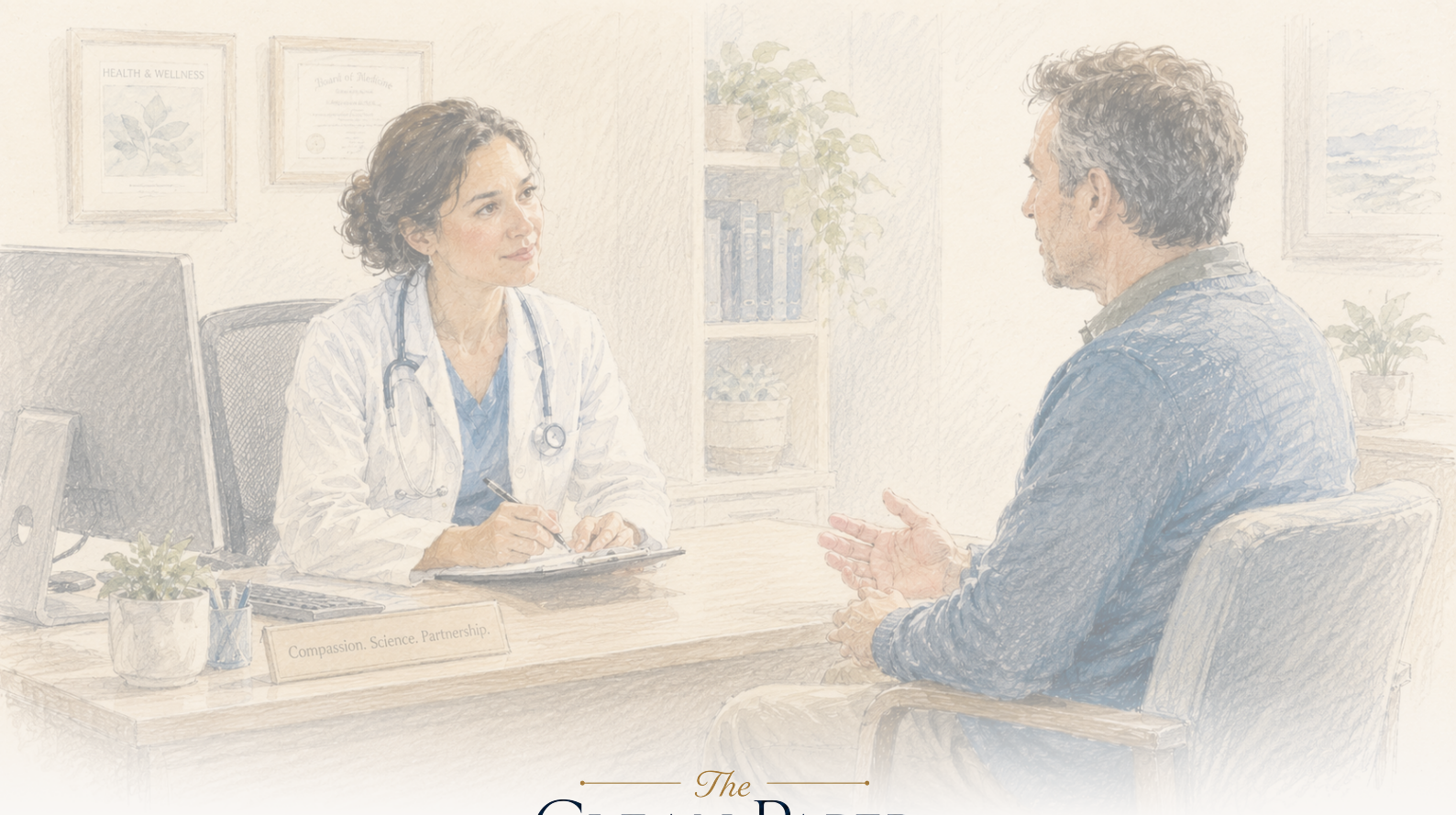


The Clean Paper

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

# Kliniczna AI wyglądała na bezpieczną i poprawiła dokumentację — lecz nie wyniki pacjentów

*Pragmatyczna, klastrowo randomizowana próba w 16 kenijskich placówkach podstawowej opieki sprawdziła generatywne narzędzie wspomagania decyzji AI dodane do rekordu clinical officers. Wśród 9691 pacjentów rygorystyczny 14-dniowy złożony wynik niepowodzenia leczenia oceniany przez ekspertów wynosił 2,2% z narzędziem i 2,0% bez niego (skorygowany iloraz szans 0,77, 95% CI 0,55–1,08, P = 0,13) — bez istotnej różnicy. Narzędzie nie wykazało sygnału zagrożenia, poprawiło dokumentację we wszystkich obszarach, nie zmieniło przepisywania ani satysfakcji i zmierzało ku niższym kosztom antybiotyków. To nie nieudane badanie, lecz rzadkie i uczciwe — mierzone na pacjentach, nie benchmarkach — pokazujące mało efektowną prawdę: narzędzie wyglądające na bezpieczne i pomagające procesowi nie przyniosło wykazanej korzyści pacjentom w dwa tygodnie, a wykrycie takiej korzyści wymagałoby znacznie większej próby.*

---

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

# Bezpieczna i pomocna nie znaczy skuteczna

Większość nagłówków o medycznej AI powstaje z niewłaściwego rodzaju badań. Model dobrze odpowiada na pytania egzaminacyjne albo pokonuje lekarzy na starannie wybranych opisach przypadków, a historia pisze się sama: maszyna jest gotowa do kliniki.

Ta próba zrobiła coś trudniejszego i rzadszego. Wprowadziła generatywne narzędzie wspomagania decyzji AI do prawdziwej podstawowej opieki zdrowotnej, w rzeczywistych placówkach i z rzeczywistymi pacjentami, po czym zadała pytanie, które naprawdę ma znaczenie: czy pacjenci uzyskali lepsze wyniki?

Uczciwa odpowiedź brzmi: nie — nie w sposób mierzalny w ciągu 14 dni. Narzędzie nie wykazało sygnału zagrożenia. Poprawiło jakość dokumentacji klinicznej. Być może nawet ograniczyło niektóre koszty leków. Nie zmniejszyło jednak istotnie liczby niepowodzeń leczenia, a autorzy ostrożnie piszą, że ewentualna korzyść jest prawdopodobnie umiarkowana.

To nie porażka badania. To prawidłowo działające badanie. Tak wygląda odpowiedzialny dowód dotyczący AI w medycynie, gdy mierzy się wyniki pacjentów, a nie benchmarki.


Narzędzie nie wykazało sygnału zagrożenia i poprawiło dokumentację kliniczną, lecz z góry określony 14-dniowy wynik pacjenta nie poprawił się istotnie. Pomoc dla procesu to nie to samo co udowodniona korzyść dla pacjenta.

# Co zrobili autorzy

Zespół przeprowadził pragmatyczne, randomizowane klastrowo badanie w **16 placówkach podstawowej opieki zdrowotnej** prywatnej sieci Penda Health w kenijskich hrabstwach Nairobi i Kiambu. Opiekę zapewniają w nich głównie **clinical officers** — pracownicy medyczni średniego szczebla z trzyletnim dyplomem — często bez łatwego dostępu do konsultacji starszego specjalisty.

Jednostką randomizacji był klinicysta, nie pacjent. Losowo przydzielono **103 clinical officers**: 52 do grupy interwencyjnej i 51 do kontrolnej. Obie korzystały z tego samego chmurowego elektronicznego rekordu medycznego. Grupa interwencyjna miała dodatkowo narzędzie **AI Consult** (wersja 2.0), oparte na dużym modelu językowym **GPT-4o firmy OpenAI** i wbudowane w rekord. Odczytywało udokumentowane przez klinicystę informacje i mogło wskazywać możliwe problemy z rozpoznaniem lub planem leczenia. Klinicyści zachowywali pełną autonomię: mogli sugestie zaakceptować, zmienić lub zignorować.

> **Jaki to był model i dlaczego szczegóły mają znaczenie**
>
> Narzędziem było **AI Consult 2.0**, działające na **GPT-4o firmy OpenAI** (wydanie z maja 2025 roku), dostępne przez komercyjne API OpenAI na licencji enterprise i ustawione na małą losowość (temperatura 0,1). Było osadzone w specjalnym elektronicznym rekordzie (EMR EasyClinic) i kierowane promptami systemowymi napisanymi pod kątem krajowych kenijskich wytycznych leczenia; autorzy opublikowali pełny prompt instrukcji.
> Dlaczego to precyzować? Ponieważ wynik dotyczy *jednego konkretnego systemu* — jednej wersji modelu, promptu, rekordu i środowiska — nie ogólnie „LLM w medycynie". Autorzy podkreślają to samo: nazywają wynik *czasowym punktem odniesienia, a nie stałym oszacowaniem zdolności*. Nowszy model, inny prompt lub słabiej scyfryzowana klinika mogą zmienić rezultat.
> Co do niezależności: OpenAI później udzieliło wsparcia rzeczowego (kredytów na obliczenia w chmurze i wskazówek technicznych dotyczących API), lecz autorzy stwierdzają, że decyzję o użyciu OpenAI podjęto *przed* tą ofertą, a firma nie uczestniczyła w projektowaniu próby, zbieraniu i analizie danych ani decyzji o publikacji.

Od 22 kwietnia do 16 lipca 2025 roku włączono **9691 pacjentów**. **Główny wynik celowo skoncentrowano na pacjencie i zdefiniowano rygorystycznie**: oceniany przez ekspertów *złożony wynik niepowodzenia leczenia* w ciągu 14 dni od wizyty — zaślepiony co do grupy panel klinicystów oceniał, czy pacjent doświadczył złego wyniku, na przykład niewyleczonej lub pogarszającej się choroby. Próbę zarejestrowano wcześniej (Pan-African Clinical Trials Registry 202502499779176).

Wybór projektu jest sednem. Łatwo wykazać, że narzędzie AI zmienia zapiski klinicysty. Znacznie trudniej — i o wiele istotniej — pokazać, że zmienia los pacjenta.

# Co odkryli

**Główny wynik nie uległ poprawie.** Niepowodzenie leczenia wystąpiło u 102 z 4693 pacjentów (2,2%) w grupie AI oraz 94 z 4654 (2,0%) w grupie kontrolnej. Surowe odsetki były minimalnie wyższe przy AI, lecz po korekcie oszacowanie punktowe skłaniało się ku korzyści: **skorygowany iloraz szans wyniósł 0,77 (95% przedział ufności 0,55–1,08, P = 0,13)** — bez istotności statystycznej. Odwrócenie kierunku między liczbami surowymi i skorygowanymi nie jest błędem rachunkowym; korekta uwzględnia różnice między klastrami klinicystów. Tak czy inaczej przedział ufności swobodnie obejmuje „brak efektu", więc nie można twierdzić, że była korzyść, a efekt bezwzględny był minimalny.

Proste objaśnienie łącznego odczytu ilorazów szans, przedziałów ufności i wartości P zawiera przewodnik po wynikach klinicznych.

**Brak sygnału zagrożenia — w pewnych granicach.** Żadnego poważnego zdarzenia niepożądanego nie uznano za związane z narzędziem, a niezależny przegląd nie znalazł sygnału bezpieczeństwa. Autorzy uczciwie wyznaczają granicę tego uspokojenia: próba nie miała mocy do wykrywania rzadkich ciężkich szkód ani z góry określonej ramy równoważności lub formalnego bezpieczeństwa, więc nie może *udowodnić* bezpieczeństwa dla rzadkich zdarzeń.

**Dokumentacja się poprawiła.** Wśród 2000 konsultacji ocenianych przez zaślepionych ekspertów klinicyści z AI Consult tworzyli lepszą dokumentację we wszystkich ocenianych dziedzinach — zapisanym rozpoznaniu, planie leczenia i ogólnej kompletności.

**Przepisywanie leków prawie się nie zmieniło.** Nie było istotnej różnicy w przepisywaniu, w tym prawidłowym użyciu antybiotyków (skorygowany iloraz szans 0,86, 95% CI 0,48–1,55). Narzędzie nie zmieniło częstości przepisywania antybiotyków.

**Pacjenci nie zauważyli różnicy.** Wśród 826 osób, które wypełniły ankietę satysfakcji, wyniki były niemal identyczne, podobnie jak czas konsultacji.

**Koszty wskazywały nieco w dół.** W analizie skorygowanej koszty związane z antybiotykami były niższe w grupie AI — prawdopodobnie dzięki tańszym wyborom, a nie mniejszej liczbie recept — i oszczędność na pacjenta wydawała się większa niż koszt działania narzędzia. Autorzy zaznaczają, że to sugestia, nie rozstrzygnięcie: pełny rachunek całkowitego kosztu posiadania wykraczał poza próbę.

Jednozdaniowe podsumowanie autorów jest najczystsze: pomoc LLM była bezpieczna w tych granicach, lecz nie ograniczyła niepowodzenia leczenia w ciągu 14 dni, a ewentualna korzyść jest prawdopodobnie umiarkowana.

# Dlaczego „brak istotnej różnicy" nie oznacza „to nie działa"

Zerowy główny wynik łatwo nadinterpretować w obie strony. Powstrzymują to dwa fakty.

Po pierwsze, **próbę zaprojektowano do wykrycia większego efektu niż zaobserwowany**. Poważne

złe wyniki są w podstawowej opiece rzadkie — tutaj około 2% — więc odróżnienie małej rzeczywistej korzyści od szumu wymaga ogromnej liczebności. Obliczenia mocy autorów wykonane po fakcie sugerują, że wykrycie efektu tej wielkości wymagałoby **znacznie większej próby, rzędu ponad 100 000 pacjentów**. Nieistotny wynik w badaniu tej wielkości nie wyklucza małej realnej korzyści; oznacza, że badanie nie potrafiło jej rozdzielić.

Po drugie, **porównanie częściowo się rozmyło**. Błąd konfiguracji na krótko dał niektórym klinicystom kontrolnym dostęp do AI Consult, a pracownicy wspólnej sieci rozmawiają ze sobą i przenoszą nawyki przez granice grup. Oba efekty upodabniają grupy i spychają każdą rzeczywistą różnicę w stronę zera. Sieć już wcześniej utrzymywała stosunkowo wysokie standardy, pozostawiając narzędziu mniej miejsca na poprawę.

Nie ratuje to nagłówka o „przełomie". Oznacza jednak, że właściwy odczyt jest skalibrowany, nie defetystyczny: w najtrudniejszym i najuczciwszym punkcie końcowym narzędzie nie poprawiło w sposób wykazany wyników pacjentów w ciągu dwóch tygodni — choć mierzalnie pomagało w dokumentacji i wyglądało bezpiecznie.

## Czego to *nie* dowodzi

- Nie pokazuje, że AI poprawiła wyniki pacjentów. Główny 14-dniowy punkt nie wykazał istotnej korzyści.
- Nie pokazuje, że AI jest bezużyteczna. Oszacowanie punktowe jej sprzyjało, dokumentacja poprawiła się we wszystkich obszarach, a koszty leków zmierzały w dół; zerowy wynik jest zgodny z małą rzeczywistą korzyścią, której próba nie mogła potwierdzić.
- Nie dowodzi bezpieczeństwa narzędzia dla rzadkich szkód. Nie wykazało sygnału zagrożenia, lecz badania nie zaprojektowano ani nie zasilono do certyfikowania rzadkich ciężkich zdarzeń.
- Nie pokazuje, że „AI pokonuje lekarzy" lub ich zastępuje. To *wspomaganie* decyzji; klinicysta zachował pełną władzę akceptowania lub odrzucania sugestii.
- Nie uogólnia się automatycznie. Próbę przeprowadzono w jednej prywatnej miejskiej sieci w Kenii; środowiska wiejskie, podmiejskie i o wyższych dochodach mogą różnić się w dowolnym kierunku.
- Nie ustala oszczędności. Sygnał kosztowy jest sugestywny, nie stanowi pełnej oceny ekonomicznej.

## Jak mocne są dowody?

Dla głównego wniosku — braku udowodnionego ograniczenia krótkoterminowej szkody pacjenta — dowód jest mocny *jako projekt* i właściwie skromny *jako konkluzja*. Prospektywna, wcześniej zarejestrowana i klastrowo randomizowana próba z zaślepionym złożonym wynikiem na poziomie pacjenta jest bliska najlepszym rzeczywistym dowodom, jakie można zebrać dla takiego narzędzia. Mówi znacznie więcej niż wynik benchmarku lub badanie opisów przypadków.

Dowody na wyniki wtórne — lepszą dokumentację, niezmienione przepisywanie, podobną satysfakcję, niższe koszty antybiotyków — są dobre, ale należy czytać je jako wtórne: wspierające sygnały, nie nagłówek, podatne na to samo zanieczyszczenie grup i ograniczenie jednej sieci.

Dowód bezpieczeństwa jest uspokajający, lecz ograniczony: nie znaleziono sygnału, ale badania nie zbudowano do wykrywania rzadkich szkód.

Najbardziej użyteczna postawa to ani „działa", ani „zawiodło". Brzmi: staranna rzeczywista próba wykazała, że narzędzie nie wzbudziło sygnału zagrożenia i poprawiło proces opieki, nie pokazując korzyści dla wyniku pacjenta w ciągu dwóch tygodni — a wykrycie takiej korzyści wymagałoby znacznie większego badania.

## Dlaczego to ważne

Debacie o medycznej AI dramatycznie brakuje właśnie takich dowodów. Tysiące prac pokazują modele zdające egzaminy i dorównujące klinicystom w uporządkowanych przypadkach. Bardzo mało jest dużych, pragmatycznych, randomizowanych prób mierzących, czy prawdziwi pacjenci mają się lepiej. To jedna z nich i prowadzi do mało efektownej prawdy: zdanie testu nie jest tym samym co pomoc pacjentowi.

Cała historia mieści się w tej luce. Narzędzie może być naprawdę użyteczne dla klinicystów — czytelniejsze notatki, niższe rachunki za leki, druga para oczu — a mimo to nie zmienić twardego wyniku pacjenta w ciągu dwóch tygodni. Oba fakty mogą być jednocześnie prawdziwe, a dojrzały system zdrowia musi zachować oba, zamiast wybierać wygodniejszy.

Praca przywraca też ciężar dowodu we właściwe miejsce. Jeśli firma chce twierdzić, że jej kliniczna AI poprawia opiekę, odpowiednim dowodem nie jest ranking. Jest nim taka próba, dotycząca wyników ważnych dla pacjentów — i najlepiej większa, ponieważ uczciwą lekcją jest to, że umiarkowane korzyści wymagają dużych badań.

## Czyste podsumowanie

Pragmatyczna, klastrowo randomizowana próba w 16 kenijskich placówkach podstawowej opieki sprawdziła generatywne narzędzie wspomagania decyzji „AI Consult" dodane do elektronicznego rekordu używanego przez clinical officers. Wśród 9691 pacjentów oceniany przez ekspertów złożony wynik niepowodzenia leczenia w ciągu 14 dni wynosił 2,2% z narzędziem i 2,0% bez niego (skorygowany iloraz szans 0,77, 95% CI 0,55–1,08, P = 0,13) — bez istotnej różnicy. Narzędzie nie wykazało sygnału zagrożenia, poprawiło dokumentację we wszystkich ocenianych obszarach, nie zmieniło przepisywania ani satysfakcji pacjentów i wiązało się z nieco niższymi kosztami antybiotyków. Autorzy wnioskują, że było bezpieczne w tych granicach, lecz nie ograniczyło niepowodzenia leczenia, a ewentualna korzyść jest prawdopodobnie umiarkowana; wykrycie efektu tej wielkości wymagałoby znacznie większej próby, rzędu 100 000 pacjentów. Wynik nie pokazuje ani że kliniczna AI poprawia

wyniki pacjentów, ani że jest bezużyteczna — pokazuje, że narzędzie wyglądające na bezpieczne i pomagające procesowi nie przyniosło wykazanej korzyści pacjentom w ciągu dwóch tygodni, w jednej prywatnej sieci miejskiej.

## Kontrola bez ściemy

**Co pokazuje praca:** W rzeczywistej randomizowanej próbie dodanie do rekordów podstawowej opieki narzędzia wspomagania decyzji opartego na LLM nie wzbudziło sygnału bezpieczeństwa, poprawiło jakość dokumentacji, nie zmieniło przepisywania i nie ograniczyło istotnie rygorystycznego 14-dniowego złożonego wyniku niepowodzenia leczenia.

**Co jest prawdopodobne, ale niedowiedzione:** Że narzędzie daje niewielkie rzeczywiste ograniczenie niepowodzeń, zbyt małe do wykrycia w tej próbie; że oszczędza pieniądze po uwzględnieniu wszystkich kosztów; że lepsza dokumentacja z czasem przekłada się na lepszą opiekę.

**Czego praca nie pokazuje:** Że kliniczna AI poprawia wyniki pacjentów; że jest bezpieczna dla rzadkich zdarzeń; że zastępuje lub przewyższa klinicystów; że wyniki przenoszą się na środowiska wiejskie lub bogatsze; że oszczędność kosztów została ustalona.

**Główne ograniczenia:** Moc zaplanowana dla większego efektu niż zaobserwowany (rzadkie wyniki potrzebują wielkich prób); błąd konfiguracji zanieczyścił grupę kontrolną, a nawyki we wspólnej sieci rozmywają porównanie — oba czynniki przesuwają w stronę braku różnicy; jedna prywatna miejska sieć o już wysokich standardach; brak z góry określonej ramy równoważności lub bezpieczeństwa; krótki 14-dniowy horyzont.

**Jak duże zaufanie powinien mieć czytelnik?** Duże do tego, że narzędzie nie wykazało sygnału zagrożenia w tej próbie i poprawiło dokumentację. Duże do tego, że nie przyniosło tutaj wykazanej poprawy 14-dniowych wyników pacjenta. Małe do dowolnego twierdzenia, że jako interwencja dla pacjentów „działa” lub „zawiodło” — pytanie naprawdę pozostaje otwarte i wymaga znacznie większej próby. Właściwa postawa: staranny rzeczywisty wynik o narzędziu, które nie wzbudziło sygnału zagrożenia i poprawiło proces opieki, nie wyrok, że AI przekształca — lub niszczy — podstawową opiekę.

## Źródła

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