The Clean Paper

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

Une IA clinique qui semblait sûre et améliorait la paperasse — mais n'a pas amélioré les résultats des patients

Un essai pragmatique, randomisé par grappes, dans 16 structures de soins primaires au Kenya a testé un outil d'aide à la décision par IA générative ajouté au dossier utilisé par les clinical officers. Chez 9 691 patients, le strict composite à 14 jours d'échec thérapeutique jugé par des experts était de 2,2 % avec l'outil contre 2,0 % sans lui (odds ratio ajusté 0,77, IC 95 % 0,55-1,08, P = 0,13) — aucune différence significative. L'outil n'a montré aucun signal de sécurité, a amélioré la documentation dans tous les domaines notés, n'a pas changé les prescriptions ni la satisfaction, et tendait vers des coûts d'antibiotiques plus bas. Ce n'est pas une étude ratée ; c'est une étude rare et honnête — mesurée sur les patients, pas sur des benchmarks — qui arrive à une vérité peu glamour : un outil qui semblait sûr et aidait le processus n'a pas démontrablement aidé les patients en deux semaines, et détecter un tel bénéfice demanderait un essai bien plus grand.

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

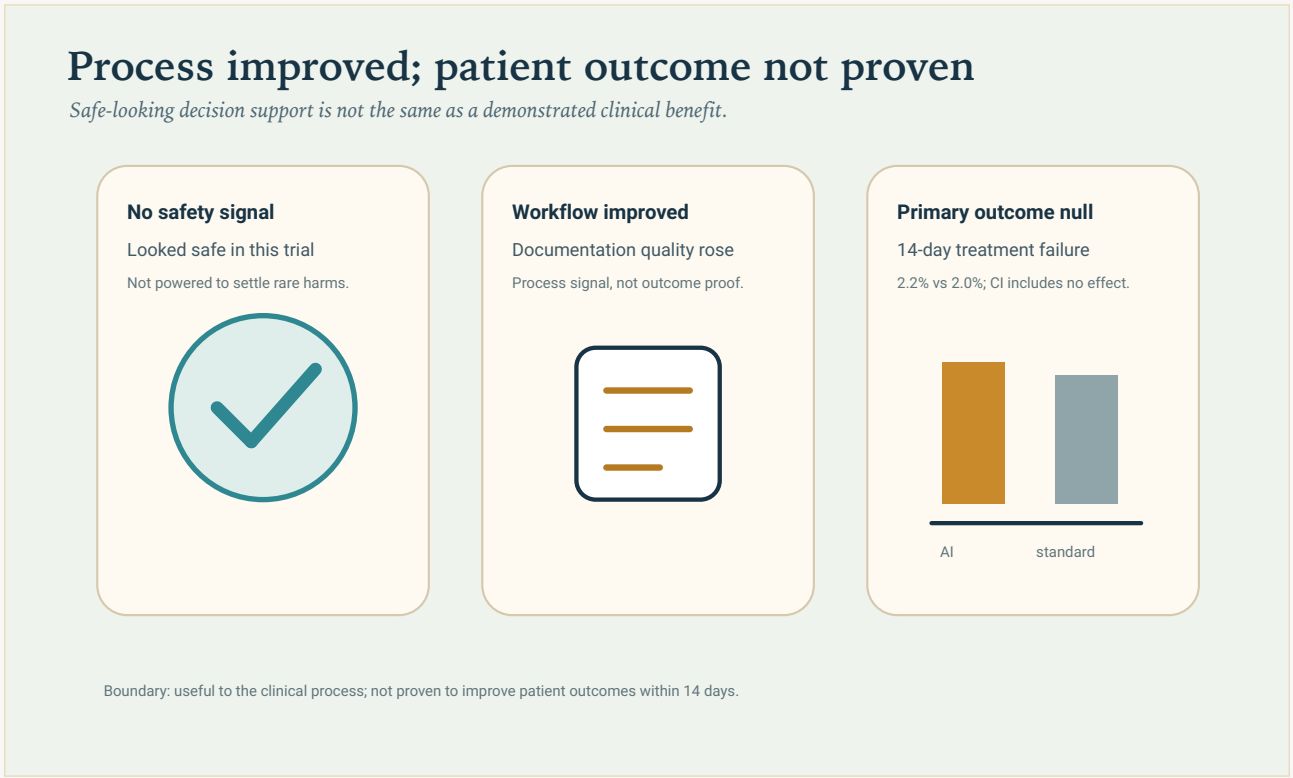
Sûr et utile n'est pas la même chose qu'efficace

La plupart des titres sur l'IA médicale sont construits à partir du mauvais type d'étude. Un modèle obtient de bons scores à des questions d'examen, ou bat des médecins sur des vignettes sélectionnées, et l'histoire s'écrit toute seule : la machine est prête pour la clinique.

Cet essai a fait quelque chose de plus difficile et plus rare. Il a mis un outil d'aide à la décision par IA générative dans de vrais soins primaires, dans de vraies structures, avec de vrais patients, et a posé la question qui compte vraiment : les patients allaient-ils mieux ?

La réponse honnête est non — pas de façon mesurable, pas en 14 jours. L'outil n'a montré aucun signal de sécurité. Il a amélioré la qualité de la documentation clinique. Il a peut-être même réduit certains coûts de médicaments. Mais il n'a pas réduit significativement les échecs thérapeutiques, et les auteurs disent prudemment que tout bénéfice, s'il existe, est probablement modeste.

Ce n'est pas un échec de l'étude. C'est l'étude qui fonctionne. Voilà à quoi ressemble une preuve responsable sur l'IA en médecine quand elle est mesurée sur des patients plutôt que sur des benchmarks.



L'outil n'a montré aucun signal de sécurité et a amélioré la documentation clinique, mais le résultat prédéfini à 14 jours pour les patients ne s'est pas significativement amélioré. Une aide au processus n'est pas la même chose qu'un bénéfice patient prouvé.

Ce que les auteurs ont fait

L'équipe a mené un essai pragmatique randomisé par grappes dans **16 structures de soins primaires** exploitées par un réseau de santé privé (Penda Health) dans les comtés de Nairobi et Kiambu, au Kenya. Les soins y sont largement fournis par des **clinical officers** — praticiens de niveau intermédiaire avec un diplôme de trois ans — souvent sans accès facile à une consultation senior.

L'unité de randomisation était le clinicien, pas le patient. **103 clinical officers** ont été randomisés : 52 dans le bras intervention et 51 dans le bras contrôle. Les deux bras utilisaient le même dossier médical électronique cloud. Le bras intervention avait en plus **« AI Consult »** (version 2.0), un outil d'aide à la décision construit sur le grand modèle de langage **GPT-4o d'OpenAI** et intégré dans ce dossier. Il lisait les informations documentées par le clinicien et pouvait signaler d'éventuels problèmes dans le diagnostic ou le plan de traitement. Les cliniciens conservaient toute leur autonomie : ils pouvaient accepter, modifier ou ignorer ses suggestions.

> **Quel modèle était-ce, et pourquoi les détails comptent**
>
> L'outil était **AI Consult 2.0**, utilisant **GPT-4o d'OpenAI** (version de mai 2025), accessible via l'API commerciale d'OpenAI sous licence entreprise et réglé avec peu d'aléa (température 0,1). Il était intégré à un dossier électronique sur mesure (l'EMR d'EasyClinic) et guidé par des prompts système écrits pour s'aligner sur les recommandations thérapeutiques nationales kényanes ; les auteurs ont publié le prompt complet.
> Pourquoi préciser tout cela ? Parce que le résultat porte sur *un système précis* — une version de modèle, un prompt, un dossier, un cadre — et non sur « les LLM en médecine » en général. Les auteurs font le même point : ils appellent leur résultat un *repère temporel plutôt qu'une estimation fixe de capacité*. Un modèle plus récent, un autre prompt ou une clinique moins numérisée pourraient tous déplacer le résultat.
> Sur l'indépendance : OpenAI a ensuite fourni un soutien en nature (crédits de calcul cloud et conseils techniques sur l'API), mais les auteurs indiquent que la décision d'utiliser OpenAI avait été prise *avant* cette offre, et qu'OpenAI n'a joué aucun rôle dans le design de l'essai, la collecte des données, l'analyse ou la décision de publier.

Entre le 22 avril et le 16 juillet 2025, **9 691 patients** ont été inclus. Le **critère principal était volontairement centré sur le patient et strict** : un *composite d'échec thérapeutique* expertisé dans les 14 jours suivant la visite — un panel de cliniciens, aveugle au bras d'étude, jugeait si chaque patient avait eu un mauvais résultat comme une maladie non résolue ou aggravée. L'essai avait été enregistré à l'avance (Pan-African Clinical Trials Registry 202502499779176).

Ce choix de design est le point central. Il est facile de montrer qu'un outil d'IA change ce qu'un clinicien écrit. Il est beaucoup plus difficile, et beaucoup plus significatif, de montrer qu'il change ce qui arrive au patient.

Ce qu'ils ont trouvé

Le critère principal ne s'est pas amélioré. Un échec thérapeutique est survenu chez 102 des 4 693 patients (2,2 %) du bras IA et 94 des 4 654 (2,0 %) du bras contrôle. Les pourcentages bruts étaient légèrement plus élevés avec l'IA, mais après ajustement l'estimation ponctuelle penchait vers un bénéfice : **l'odds ratio ajusté était de 0,77 (intervalle de confiance à 95 % 0,55 à 1,08, P = 0,13)** — non statistiquement significatif. Ce basculement entre les nombres bruts et ajustés n'est pas une erreur arithmétique ; l'ajustement tient compte des différences entre grappes de cliniciens. Dans tous les cas, l'intervalle de confiance inclut largement « pas d'effet », donc aucun bénéfice ne peut être revendiqué, et en termes absolus l'effet était minuscule.

Pour une explication simple des odds ratios, intervalles de confiance et valeurs P ensemble, voir le guide de lecture d'un résultat clinique.

Aucun signal de sécurité — dans certaines limites. Aucun événement indésirable grave n'a été jugé lié à l'outil, et une revue indépendante n'a trouvé aucun signal de sécurité. Les auteurs sont honnêtes sur le plafond de cette reassurance : l'essai n'avait pas la puissance pour détecter des dommages graves rares, et il n'avait pas de cadre formel de non-infériorité ou de sécurité prédéfini, donc il ne peut pas *prouver* la sécurité pour des événements peu fréquents.

La documentation s'est améliorée. Parmi 2 000 consultations relues par des experts aveugles, les cliniciens utilisant AI Consult ont produit une meilleure documentation clinique dans tous les domaines évalués : diagnostic noté, plan de traitement et complétude générale.

Les prescriptions ont à peine bougé. Il n'y avait pas de différence significative dans les prescriptions, y compris l'usage correct des antibiotiques (odds ratio ajusté 0,86, IC 95 % 0,48 à 1,55). L'outil n'a pas changé les taux de prescription d'antibiotiques.

Les patients n'ont pas perçu de différence. Parmi 826 patients ayant rempli une enquête de satisfaction, la satisfaction était essentiellement identique entre les bras, et les temps de consultation étaient similaires.

Les coûts pointaient légèrement vers le bas. Dans une analyse ajustée, les coûts liés aux antibiotiques étaient plus bas dans le bras IA — plausiblement par des choix moins chers plutôt que par moins de prescriptions — et l'économie par patient sur les antibiotiques semblait dépasser le coût par patient de fonctionnement de l'outil. Les auteurs le signalent comme suggestif, pas réglé : une comptabilité complète du coût total de possession était hors du champ de l'essai.

La phrase de résumé des auteurs est la version la plus propre : l'aide par LLM était sûre dans ces limites mais n'a pas réduit l'échec thérapeutique en 14 jours, et tout bénéfice est probablement modeste.

Pourquoi « pas de différence significative » ne veut pas dire « ça ne marche pas »

Un critère principal nul est facile à surlire dans les deux sens. Deux choses empêchent l'histoire simple.

D'abord, **l'essai était construit pour détecter un effet plus grand que celui observé**. Les mauvais résultats graves en soins primaires sont rares — environ 2 % ici — donc distinguer un petit bénéfice réel du bruit demande des nombres énormes. Les calculs de puissance post-hoc des auteurs suggèrent que détecter un effet de la taille observée demanderait **un essai beaucoup plus grand, de l'ordre de plus de 100 000 patients**. Un résultat non significatif dans une étude de cette taille n'exclut pas un petit bénéfice réel ; il signifie que cette étude ne pouvait pas le résoudre.

Ensuite, **la comparaison était partiellement brouillée**. Une erreur de configuration a brièvement donné accès à AI Consult à certains cliniciens du bras contrôle, et les cliniciens d'un réseau partagé se parlent et transportent des habitudes d'un côté à l'autre. Les deux effets tendent à rendre les deux bras plus semblables, poussant toute différence réelle vers zéro. En plus, le réseau hôte fonctionnait déjà avec des standards relativement élevés, ce qui laisse moins de marge à un outil pour montrer une amélioration.

Rien de cela ne sauve un titre de « percée ». Mais cela signifie que la bonne lecture est calibrée, pas déflationniste : sur le critère le plus dur et le plus honnête, cet outil n'a pas démontrablement aidé les patients en deux semaines — tout en aidant mesurablement la tenue du dossier et en semblant sûr.

Ce que cela ne prouve *pas*

- Cela ne montre **pas** que l'IA a amélioré les résultats des patients. Sur le critère principal à 14 jours, il n'y avait pas de bénéfice significatif.
- Cela ne montre **pas** que l'IA est inutile. L'estimation ponctuelle la favorisait, la documentation s'est améliorée partout et les coûts de médicaments tendaient à baisser ; le résultat nul est compatible avec un petit bénéfice réel que l'essai était trop petit pour confirmer.
- Cela ne prouve **pas** que l'outil est sûr pour les dommages rares. Il n'a montré aucun signal de sécurité, mais il n'était pas assez puissant ni conçu pour certifier la sécurité pour des événements graves peu fréquents.
- Cela ne montre **pas** que « l'IA bat les médecins » ni remplace les cliniciens. C'est une aide à la décision ; le clinicien gardait toute autorité pour l'accepter ou la rejeter.
- Cela ne se généralise **pas** automatiquement. L'essai s'est déroulé dans un seul réseau privé urbain au Kenya ; les contextes ruraux, périurbains ou à revenu plus élevé pourraient différer dans les deux sens.
- Cela n'établit **pas** des économies de coûts. Le signal de coût est suggestif, pas une évaluation économique complète.

Quelle est la force des preuves ?

Pour l'affirmation centrale — pas de réduction prouvée du dommage patient à court terme — la preuve est forte *comme design* et justement humble *comme conclusion*. Un essai prospectif, pré-enregistré, randomisé par grappes, avec un critère composite patient aveuglé, est proche de la meilleure preuve réelle qu'on puisse rassembler pour un tel outil. C'est beaucoup plus informatif qu'un score de benchmark ou une étude de vignettes.

Pour les résultats secondaires — meilleure documentation, prescriptions inchangées, satisfaction similaire, coûts d'antibiotiques plus bas — la preuve est bonne mais doit être lue comme secondaire : des signaux de soutien, pas le titre principal, et vulnérables aux mêmes limites de contamination et de réseau unique.

Pour la sécurité, la preuve est rassurante mais bornée : aucun signal trouvé, mais pas une étude conçue pour trouver des dommages rares.

La posture la plus utile n'est ni « ça marche » ni « ça a échoué ». Elle est : un essai prudent en conditions réelles a trouvé que cet outil d'IA ne soulevait aucun signal de sécurité et améliorait le processus de soins, sans démontrer de bénéfice sur les résultats patients en deux semaines — et que détecter un tel bénéfice demanderait une étude beaucoup plus grande.

Pourquoi cela compte

Le débat sur l'IA médicale manque cruellement de ce type de preuve. Il existe des milliers d'articles montrant des modèles réussir des examens et égaler des cliniciens sur des cas bien propres. Il y a très peu de grands essais pragmatiques randomisés mesurant si de vrais patients vont mieux. Celui-ci en fait partie, et il tombe sur une vérité peu glamour : réussir le test n'est pas la même chose qu'aider le patient.

Cet écart est toute l'histoire. Un outil peut être réellement utile aux cliniciens — notes plus claires, factures de médicaments plus basses, deuxième paire d'yeux — et ne pas déplacer un critère dur de résultat patient en quinze jours. Les deux faits peuvent être vrais en même temps, et un système de santé mature doit les tenir ensemble plutôt que choisir celui qui l'arrange.

Cela replace aussi la charge de la preuve dans une direction utile. Si une entreprise veut affirmer que son IA clinique améliore les soins, la preuve pertinente n'est pas un classement. C'est un essai comme celui-ci, sur des résultats qui comptent pour les patients — et idéalement un plus grand, car la leçon honnête ici est que les bénéfices modestes demandent de grandes études pour être vus.

Résumé propre

Un essai pragmatique randomisé par grappes dans 16 structures de soins primaires au Kenya a testé un outil d'aide à la décision par IA générative (« AI Consult ») ajouté au dossier électronique utilisé par des clinical officers. Chez 9 691 patients, le composite expertisé d'échec thérapeutique dans les 14 jours était de 2,2 % avec l'outil contre 2,0 % sans lui (odds ratio ajusté 0,77, IC 95 % 0,55-1,08, P = 0,13) — aucune différence significative. L'outil n'a montré aucun signal de sécurité, a amélioré la documentation clinique dans tous les domaines notés, n'a pas changé les prescriptions, a laissé la satisfaction des patients inchangée et était associé à des coûts d'antibiotiques un peu plus bas. Les auteurs concluent qu'il était sûr dans ces limites mais n'a pas réduit l'échec thérapeutique, avec un bénéfice probablement modeste ; détecter un effet de la taille observée demanderait un essai beaucoup plus grand (de l'ordre de 100 000 patients). Le résultat ne montre pas que l'IA clinique améliore les résultats patients, ni qu'elle est inutile — il montre qu'un outil qui semblait sûr et aidait le processus n'a pas démontrablement aidé les patients en deux semaines, dans un réseau privé urbain.

No-BS check

Ce que l'article montre : Dans un essai randomisé en conditions réelles, l'ajout d'un outil d'aide à la décision fondé sur un LLM aux dossiers de soins primaires n'a soulevé aucun signal de sécurité, a amélioré la qualité de la documentation clinique, n'a pas changé les prescriptions et n'a pas réduit significativement un composite strict d'échec thérapeutique à 14 jours.

Ce qui est plausible mais non prouvé : Que l'outil produise une petite réduction réelle de l'échec thérapeutique trop petite pour que cet essai la détecte ; qu'il économise de l'argent une fois tous les coûts comptés ; qu'une meilleure documentation finisse par se traduire en meilleurs soins.

Ce que cela ne montre pas : Que l'IA clinique améliore les résultats patients ; qu'elle est sûre pour des événements rares ; qu'elle remplace ou surpasse les cliniciens ; que ces résultats se transfèrent aux milieux ruraux ou à revenu plus élevé ; que l'économie de coûts est établie.

Principales limites : Puissance prévue pour un effet plus grand que celui observé (les événements rares exigent de très grands échantillons) ; une erreur de configuration a contaminé le bras contrôle et les habitudes d'un réseau partagé brouillent la comparaison, deux biais vers l'absence de différence ; réseau urbain privé unique avec des standards déjà élevés ; pas de cadre formel de non-infériorité ou de sécurité prédéfini ; horizon court de 14 jours.

Quel niveau de confiance pour un lecteur généraliste ? Élevé que l'outil n'ait montré aucun signal de sécurité dans cet essai et ait amélioré la documentation. Élevé qu'il n'ait pas démontrablement amélioré ici les résultats patients à 14 jours. Bas pour toute affirmation qu'il « marche » ou « échoue » comme intervention de bénéfice patient — la question reste réellement non résolue et demande une étude bien plus grande. Position appropriée : un résultat prudent en conditions réelles sur un outil qui n'a pas soulevé de signal de sécurité et a amélioré le processus de soins, pas un verdict que l'IA transforme — ou détruit — les soins primaires.

Sources

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