The Clean Paper

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

안전해 보였고 문서 작업은 개선했지만 환자 결과는 개선하지 못한 임상 AI

케냐의 *16* 개 일차의료 시설에서 진행된 실용적 군집 무작위배정 임상시험은 임상 담당관들이 사용하는 기록에 생성형 *AI* 의사결정 지원 도구를 추가해 평가했다. *9,691* 명의 환자에서 전문가가 판정한 엄격한 *14* 일 치료 실패 복합 결과는 도구 사용 시 *2.2%*, 미사용 시 *2.0%* 였다 *(*조정 오즈비 *0.77, 95% CI 0.55-1.08, P = 0.13)*. 유의한 차이는 없었다. 이 도구에서는 안전성 신호가 나타나지 않았고, 평가된 모든 영역에서 문서화가 개선됐으며, 처방과 만족도는 변하지 않았고, 항생제 비용은 낮아지는 추세를 보였다. 이는 실패한 연구가 아니다. 벤치마크가 아니라 환자를 대상으로 측정한 드물고 정직한 연구다. 안전해 보였고 진료 과정을 도운 도구가 *2* 주 안에 환자에게 명확한 도움을 주지는 못했으며, 그런 편익이 있다면 이를 발견하려면 훨씬 더 큰 임상시험이 필요하다는, 화려하지 않은 진실에 도달한 연구다.

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

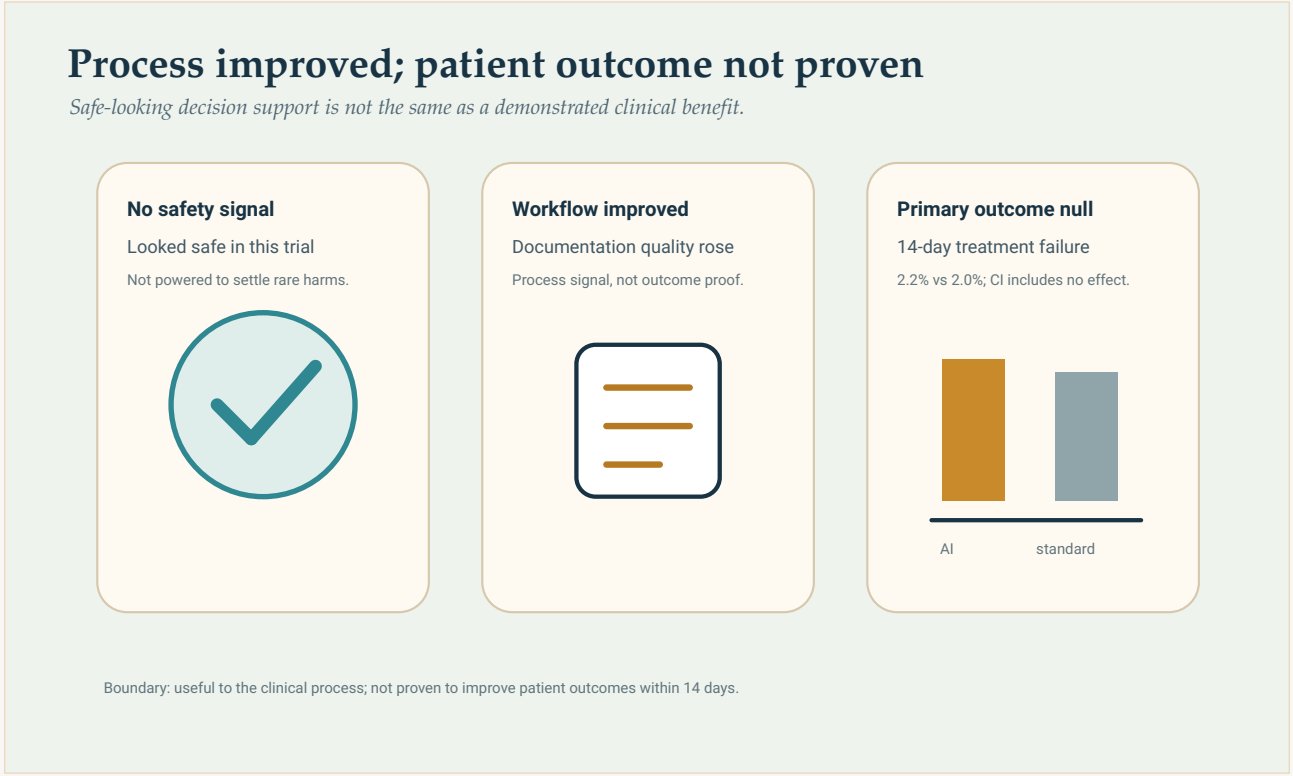
안전하고 도움이 된다고 해서 효과적인 것은 아니다

의료 AI 에 관한 헤드라인 대부분은 잘못된 종류의 연구에 기반한다. 어떤 모델이 시험에서 높은 점수를 받거나, 선별된 증례에서 의사를 앞서면 이야기는 저절로 완성된다. 기계가 진료 현장에 나설 준비가 됐다는 것이다.

이 임상시험은 더 어렵고 드문 일을 했다. 생성형 AI 의사결정 지원 도구를 실제 일차의료, 실제 시설, 실제 환자에게 적용하고, 정말 중요한 질문을 던졌다. 환자들이 더 나아졌는가?

솔직한 답은 아니다. 적어도 측정 가능한 수준에서는, 14 일 안에는 그렇지 않았다. 이 도구와 관련된 안전성 신호는 나타나지 않았다. 임상 문서의 질은 개선됐다. 일부 약제 비용을 낮췄을 가능성도 있다. 그러나 치료 실패를 유의하게 줄이지는 못했고, 저자들은 효과가 존재하더라도 아마 크지 않을 것이라고 신중하게 말한다.

이는 연구의 실패가 아니다. 연구가 제대로 작동한 것이다. AI 를 벤치마크가 아니라 환자를 대상으로 측정했을 때, 의학에서 책임 있는 AI 근거란 이런 모습이다.



이 도구와 관련된 안전성 신호는 나타나지 않았고 임상 문서화는 개선됐지만, 사전에 정한 14 일 환자 결과는 유의하게 개선되지 않았다. 진료 과정에 도움이 된다는 것과 입증된 환자 편익은 같은 말이 아니다.

저자들이 한 일

연구팀은 케냐 나이로비와 키암부 카운티에서 민간 의료 네트워크 (Penda Health) 가 운영하는 **16 개 일차의료 시설**을 대상으로 실용적 군집 무작위배정 임상시험을 실시했다. 이 시설에서는 주로 **임상진료관 (clinical officers)** 이 진료를 제공한다. 이들은 삼 년 과정의 디플로마를 취득한 중간 수준의 의료인으로, 상급자에게 쉽게 자문을 구하기 어려운 경우가 많다.

무작위배정 단위는 환자가 아니라 임상의였다. **103 명의 임상진료관**을 무작위배정해 52 명은 중재군, 51 명은 대조군에 배정했다. 두 군 모두 동일한 클라우드 기반 전자의무기록을 사용했다. 중재군에는 여기에 더해 **"AI Consult"**(버전 2.0) 가 제공됐다. 이는 **OpenAI 의 GPT-4o** 대규모 언어 모델을 기반으로 구축된 의사결정 지원 도구로, 전자의무기록에 내장돼 있었다. 이 도구는 임상의가 기록한 정보를 읽고 진단이나 치료 계획에 있을 수 있는 문제를 표시할 수 있었다. 임상의의 자율성은 전적으로 유지됐다. 제안을 수용하거나 수정하거나 무시할 수 있었다.

> **어떤 모델이었으며, 구체적인 세부사항이 중요한 이유**
>
> 이 도구는 **AI Consult 2.0** 으로, **OpenAI 의 GPT-4o**(**2025 년 5 월** 출시 버전) 를 사용했다. 기업용 라이선스에 따른 OpenAI 의 상용 API 를 통해 접근했으며, 무작위성을 낮춘 설정 (temperature 0.1) 에서 실행됐다. 이 도구는 맞춤형 전자의무기록 (EasyClinic 의 EMR) 안에 들어갔고, 케냐 국가 치료 지침에 맞도록 작성된 시스템 프롬프트의 지시를 받았다. 저자들은 전체 지시 프롬프트를 공개했다.
> 왜 이를 구체적으로 밝힐까? 이 결과는 "의학에서의 LLM" 전반에 관한 것이 아니라, 하나의 특정 시스템—하나의 모델 버전, 하나의 프롬프트, 하나의 기록 시스템, 하나의 환경—에 관한 것이기 때문이다. 저자들도 같은 점을 지적한다. 이들은 자신들의 발견을 능력에 대한 고정된 추정치가 아니라 시간적 벤치마크라고 부른다. 더 새로운 모델, 다른 프롬프트, 또는 디지털화 수준이 낮은 진료소라면 결과가 어느 방향으로든 달라질 수 있다.
> 독립성에 관해서는, OpenAI 가 이후 현물 지원 (클라우드 컴퓨팅 크레딧과 API 사용에 관한 기술적 안내) 을 제공했지만, 저자들은 OpenAI 를 사용하기로 한 결정이 그 제안보다 앞서 내려졌으며 OpenAI 는 임상시험의 설계, 자료 수집, 분석 또는 출판 결정에 관여하지 않았다고 밝혔다.

2025 년 4 월 22 일부터 7 월 16 일까지 **9,691 명의 환자**가 등록됐다. **주요 결과는 의도적으로 환자 중심이며 엄격하게 설정됐다.** 방문 후 14 일 이내의 전문가 판정 치료 실패 복합 결과였다. 즉, 연구군을 알지 못하는 임상의 패널이 각 환자에게 호전되지 않거나 악화되는 질환과 같은 나쁜 결과가 있었는지를 판정했다. 이 임상시험은 사전에 등록됐다 (범아프리카 임상시험 등록소 202502499779176).

이 설계 선택이 핵심이다. AI 도구가 임상의의 기록 내용을 바꾼다는 점을 보여주기는 쉽다. 환자에게 실제로 일어나는 일을 바꾼다는 점을 보여주는 것은 훨씬 어렵고, 훨씬 더 의미 있다.

발견한 내용

주요 결과는 개선되지 않았다. 치료 실패는 AI 군에서 4,693 명 중 102 명 (2.2%), 대조군에서 4,654 명 중 94 명 (2.0%) 에게 발생했다. 조정하지 않은 비율은 AI 사용군에서 아주 조금 높았지만, 조정 후 점추정치

는 편익을 시사했다. **조정 오즈비는 0.77(95% 신뢰구간 0.55–1.08, P = 0.13)** 로, 통계적으로 유의하지 않았다. 원자료와 조정 후 수치가 달라진 것은 산술 오류가 아니다. 조정은 임상의 군집 간 차이를 반영한다. 어느 쪽으로 보더라도 신뢰구간에는 "효과 없음" 이 충분히 포함되므로 편익을 주장할 수 없으며, 절대적인 측면에서 효과도 매우 작았다.

오즈비, 신뢰구간, P 값을 함께 읽는 방법을 쉬운 말로 설명한 내용은 임상 결과 읽기 안내서를 참고하면 된다.

안전성 신호는 없었다—단, 한계가 있다. 이 도구와 관련 있다고 판단된 중대한 이상반응은 없었고, 독립적인 검토에서도 안전성 신호는 발견되지 않았다. 저자들은 이러한 안심의 한계를 솔직하게 밝힌다. 이 임상시험은 드물고 심각한 위해를 발견하도록 설계된 검정력을 갖추지 않았고, 사전에 정한 비열등성 또는 공식적인 안전성 평가 체계도 없었다. 따라서 드문 사건에 대한 안전성을 입증할 수는 없다.

문서화는 개선됐다. 눈가림된 전문가들이 검토한 2,000 건의 진료 중 AI Consult 를 사용한 임상의들은 평가된 모든 영역에서 더 나은 임상 문서를 작성했다. 기록된 진단, 치료 계획, 전반적인 완전성이 모두 개선됐다.

처방은 거의 변하지 않았다. 올바른 항생제 사용을 포함한 처방에서 유의한 차이는 없었다 (조정 오즈비 0.86, 95% CI 0.48 to 1.55). 이 도구는 항생제 처방률을 바꾸지 않았다.

환자들은 차이를 느끼지 못했다. 만족도 설문을 완료한 826 명의 환자에서 만족도는 두 군 사이에 사실상 동일했고, 진료 시간도 비슷했다.

비용은 약간 낮아지는 방향을 보였다. 조정 분석에서 항생제 관련 비용은 AI 군에서 더 낮았다. 이는 처방 횟수가 줄어서라기보다 더 저렴한 선택을 했기 때문일 가능성이 있다. 환자 한 명당 항생제 절감액은 도구 운영의 환자 한 명당 비용을 웃도는 것으로 나타났다. 저자들은 이를 확정된 결과가 아니라 시사적인 신호로 제시한다. 총소유비용을 완전히 산정하는 일은 이 임상시험의 범위를 벗어났다.

저자들이 제시한 한 문장 요약이 가장 깔끔하다. LLM 지원은 이러한 한계 안에서 안전했지만 14 일 이내의 치료 실패를 줄이지 못했으며, 어떤 편익이 있더라도 아마 크지 않을 것이다.

"유의한 차이 없음" 은 "효과가 없다" 는 뜻이 아니다

주요 결과에서 유의한 차이가 없었다는 사실은 어느 방향으로든 쉽게 과도하게 해석될 수 있다. 단순한 해석을 막는 요소는 두 가지다.

첫째, **이 임상시험은 관찰된 효과보다 더 큰 효과를 포착하도록 설계됐다.** 일차의료에서 심각한 나쁜 결과는 드물다. 여기서는 약 2% 였다. 따라서 작지만 실제인 편익과 잡음을 구분하려면 엄청난 표본이 필요하다. 저자들의 사후 검정력 계산에 따르면, 관찰된 정도의 효과를 발견하려면 **훨씬 더 큰 임상시험, 100,000 명을 넘는 규모가 필요할 수 있다.** 이 정도 규모의 연구에서 통계적으로 유의하지 않은 결과가 나왔다고 해서 작지만 실제인 편익을 배제할 수는 없다. 이 연구가 그런 편익을 판별할 수 없었다는 뜻이다.

둘째, **비교가 일부 흐려졌다.** 설정 오류로 인해 잠시 일부 대조군 임상의들이 AI Consult 에 접근할 수 있었고, 같은 네트워크에 속한 임상의들은 서로 대화하며 경계를 넘어 습관을 공유한다. 두 효과 모두 두 군을 더 비슷하게 보이게 만들어 실제 차이를 0 에 가깝게 보이게 만들 수 있다. 여기에 더해 해당 의료 네트워크는 이미 비교적 높은 수준의 진료를 제공하고 있었으므로, 도구가 개선을 보여줄 여지가 더 적었다.

그렇다고 "획기적" 이라는 헤드라인이 되살아나는 것은 아니다. 그러나 올바른 해석은 평가절하가 아니라 균형 잡힌 해석이어야 한다. 가장 어렵고 정직한 결과 지표에서 이 도구는 두 주 안에 환자에게 명확한 편익을 보여주지 못했지만, 기록 작성에는 측정 가능한 도움을 주었고 안전해 보였다.

이것이 입증하지 않는 것

- AI 가 환자 결과를 개선했다는 것을 **보여주지 않는다**. 주요 14 일 결과 지표에서 유의한 편익은 없었다.
- AI 가 쓸모없다는 것을 **보여주지 않는다**. 점추정치는 AI 에 유리했고, 문서화는 전 영역에서 개선됐으며, 약제 비용은 하락하는 추세를 보였다. 영 결과는 임상시험이 확인하기에는 너무 작았던 실제 편익과 양립할 수 있다.
- 드문 위해에 대해 도구가 안전하다는 것을 **입증하지 않는다**. 안전성 신호는 나타나지 않았지만, 드물고 심각한 사건에 대한 안전성을 인증할 검정력이나 설계를 갖춘 연구는 아니었다.
- "AI 가 의사를 이긴다" 거나 임상의를 대체한다는 것을 **보여주지 않는다**. 이는 의사결정 지원 도구이며, 임상의는 이를 받아들이거나 거부할 전적인 권한을 유지했다.
- 결과를 자동으로 일반화할 수 있다는 것을 **보여주지 않는다**. 이 임상시험은 케냐의 한 민간 도시 네트워크에서 수행됐다. 농촌, 도시 주변부 및 고소득 환경에서는 결과가 어느 방향으로든 달라질 수 있다.
- 비용 절감을 확립하지 않는다. 비용 신호는 시사적일 뿐, 완결된 경제성 평가가 아니다.

근거는 얼마나 탄탄한가?

핵심 주장, 즉 단기적인 환자 위해의 입증된 감소가 없었다는 주장에 대해 근거는 설계 측면에서는 강하고, 결론은 그에 맞게 겸손하다. 전향적이고 사전 등록됐으며 군집 무작위배정으로 진행된 임상시험에서, 눈가림된 환자 수준의 복합 결과를 사용했다는 점은 이와 같은 도구에 대해 실제 환경에서 수집할 수 있는 최선의 근거에 가깝다. 이는 벤치마크 점수나 증례 연구보다 훨씬 많은 정보를 제공한다.

이차 결과인 문서화 개선, 처방 변화 없음, 비슷한 만족도, 항생제 비용 감소에 대해서는 근거가 양호하지만 이차 결과로 읽어야 한다. 즉, 지지하는 신호이지 핵심 결론이 아니며, 동일한 오염과 단일 네트워크라는 한계의 영향을 받을 수 있다.

안전성에 대해서는 근거가 안심할 만하지만 범위가 제한적이다. 신호는 발견되지 않았지만, 드문 위해를 찾도록 설계된 연구는 아니었다.

가장 유용한 태도는 "효과가 있다" 도 "실패했다" 도 아니다. 이렇게 정리할 수 있다. 신중하게 설계된 실제 환경 임상시험에서 이 AI 도구는 안전성 신호를 높이지 않았고 진료 과정을 개선했지만, 두 주 안에 환자 결과의 편익을 입증하지는 못했다. 그리고 그러한 편익이 있다면 이를 발견하려면 훨씬 더 큰 연구가 필요하다.

중요한 이유

의료 AI 에 관한 논쟁에는 바로 이런 종류의 근거가 극도로 부족하다. 모델이 시험에서 뛰어난 성적을 거두고 정돈된 증례에서 임상의와 비슷한 성과를 냈다는 논문은 수천 편이다. 실제 환자들이 더 나아졌는지를 측정한 대규모 실용적 무작위 임상시험은 극히 드물다. 이 연구는 그중 하나이며, 화려하지 않은 진실에 도달한다. 시험을 통과하는 것과 환자를 돕는 것은 같은 말이 아니다.

이 간극이 이야기의 전부다. 도구가 임상의에게 실제로 유용할 수 있다. 더 명확한 기록, 더 낮은 약제 비용, 또 하나의 검토 관점이 그 예다. 그러면서도 두 주라는 기간 안에 중요한 환자 결과를 바꾸지 못할 수 있다. 두 사실은 동시에 참일 수 있으며, 성숙한 보건의료 체계는 편리한 한쪽을 고르는 대신 두 사실을 함께 다뤄야 한다.

이는 입증 책임의 방향도 유용하게 바꾼다. 어떤 기업이 자신의 임상 AI 가 진료를 개선한다고 주장하려면 관련 근거는 순위표가 아니다. 환자에게 중요한 결과를 대상으로 한 이와 같은 임상시험이며, 이상적으로는 더 큰 규모의 연구여야 한다. 여기서 얻는 정직한 교훈은 작은 편익을 확인하려면 대규모 연구가 필요하다는 것이기 때문이다.

핵심 요약

케냐의 **16 개 일차의료 시설**에서 실시된 실용적 군집 무작위배정 임상시험은 임상진료관들이 사용하는 전자의무기록에 생성형 AI 의사결정 지원 도구 (“AI Consult”) 를 추가해 평가했다. 9,691 명의 환자에서 14 일 이내 전문가가 판정한 치료 실패 복합 결과는 도구 사용 시 2.2%, 미사용 시 2.0% 였다 (조정 오즈비 0.77, 95% CI 0.55–1.08, P = 0.13). 유의한 차이는 없었다. 이 도구와 관련된 안전성 신호는 나타나지 않았고, 평가된 모든 영역에서 임상 문서화가 개선됐으며, 처방은 변하지 않았고, 환자 만족도도 그대로였으며, 항생제 비용은 다소 낮아지는 양상을 보였다. 저자들은 이러한 한계 안에서 도구가 안전했지만 치료 실패를 줄이지는 못했고, 편익이 있다면 아마 크지 않을 것이라고 결론지었다. 관찰된 정도의 효과를 발견하려면 훨씬 더 큰 임상시험 (100,000 명 규모) 이 필요할 것이다. 이 결과는 임상 AI 가 환자 결과를 개선한다는 것을 보여주지도, 쓸모없다는 것을 보여주지도 않는다. 안전해 보였고 진료 과정을 도왔던 도구가 한 도시의 민간 네트워크에서 두 주 안에 환자에게 명확한 도움을 주지는 못했다는 것을 보여줄 뿐이다.

과장 없이 보면

논문이 보여주는 것: 실제 환경에서 무작위배정으로 진행된 임상시험에서 LLM 기반 의사결정 지원 도구를 일차의료 기록에 추가하자 안전성 신호는 나타나지 않았고, 임상 문서의 질은 개선됐으며, 처방은 변하지 않았고, 엄격한 14 일 환자 치료 실패 복합 결과도 유의하게 줄지 않았다.

가능하지만 입증되지 않은 것: 이 도구가 이 임상시험에서 발견하기에는 너무 작은 실제 치료 실패 감소를 만들어낼 가능성, 전체 비용을 계산한 뒤에도 비용을 절감할 가능성, 더 나은 문서화가 결국 더 나은 진료로 이어질 가능성.

보여주지 않는 것: 임상 AI 가 환자 결과를 개선한다는 것, 드문 사건에 대해 안전하지 않다는 것, 임상의를

대체하거나 능가한다는 것, 이러한 결과가 농촌 또는 고소득 환경에도 적용된다는 것, 비용 절감이 확립됐다는 것.

주요 한계: 관찰된 것보다 더 큰 효과를 탐지할 수 있도록 검정력이 설계됨 (드문 결과에는 매우 큰 표본이 필요함); 설정 오류로 대조군이 오염됐고 공유 네트워크의 습관이 비교를 흐렸으며, 두 요인 모두 차이 없음 쪽으로 편향시킴; 이미 높은 기준을 갖춘 하나의 민간 도시 네트워크에서 수행됨; 사전에 정한 비열등성 또는 안전성 평가 체계가 없음; 짧은 14 일 관찰 기간.

일반 독자는 어느 정도 확신을 가져야 하는가? 이 임상시험에서 도구가 안전성 신호를 보이지 않았고 문서화를 개선했다는 점에 대해서는 높은 확신을 가져도 된다. 이 환경에서 14 일 환자 결과를 명확하게 개선하지 못했다는 점에 대해서도 높은 확신을 가질 수 있다. 환자 편익을 위한 개입으로서 이 도구가 "효과가 있다" 거나 "실패했다" 고 주장하는 데 대한 확신은 낮아야 한다. 그 질문은 실제로 미해결 상태이며 훨씬 더 큰 연구가 필요하다. 적절한 태도는 다음과 같다. 이는 안전성 신호를 높이지 않았고 진료 과정을 개선한 도구에 관한 신중한 실제 환경 결과이지, AI 가 일차의료를 혁신한다거나 망친다는 판결이 아니다.

출처

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