The Clean Paper

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

一个临床 AI 看起来安全且改善了书面流程——但并未改善患者结局

一项在 *16* 家肯尼亚初级医疗设施中进行的实用性集群随机试验，测试了一个嵌入临床官员所用电子病历的生成式 *AI* 决策支持工具。在 *9,691* 名患者中，专家裁定的 *14* 天严格治疗失败复合指标在使用该工具时为 *2.2%*，未使用时为 *2.0%*（调整后比值比 *0.77*，*95% CI 0.55-1.08*，*P = 0.13*）——无显著差异。该工具未显示安全信号，改善了所有评分领域的文档质量，使处方和患者满意度不变，并呈抗生素成本下降趋势。这不是一项失败的研究；它是一项罕见而诚实的研究——用患者而非基准测试来衡量——结果指向一个不光彩的事实：一个看起来安全且帮助了流程的工具，在两周内未能明显帮助患者，而要检测任何此类益处将需要一项大得多的试验。

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

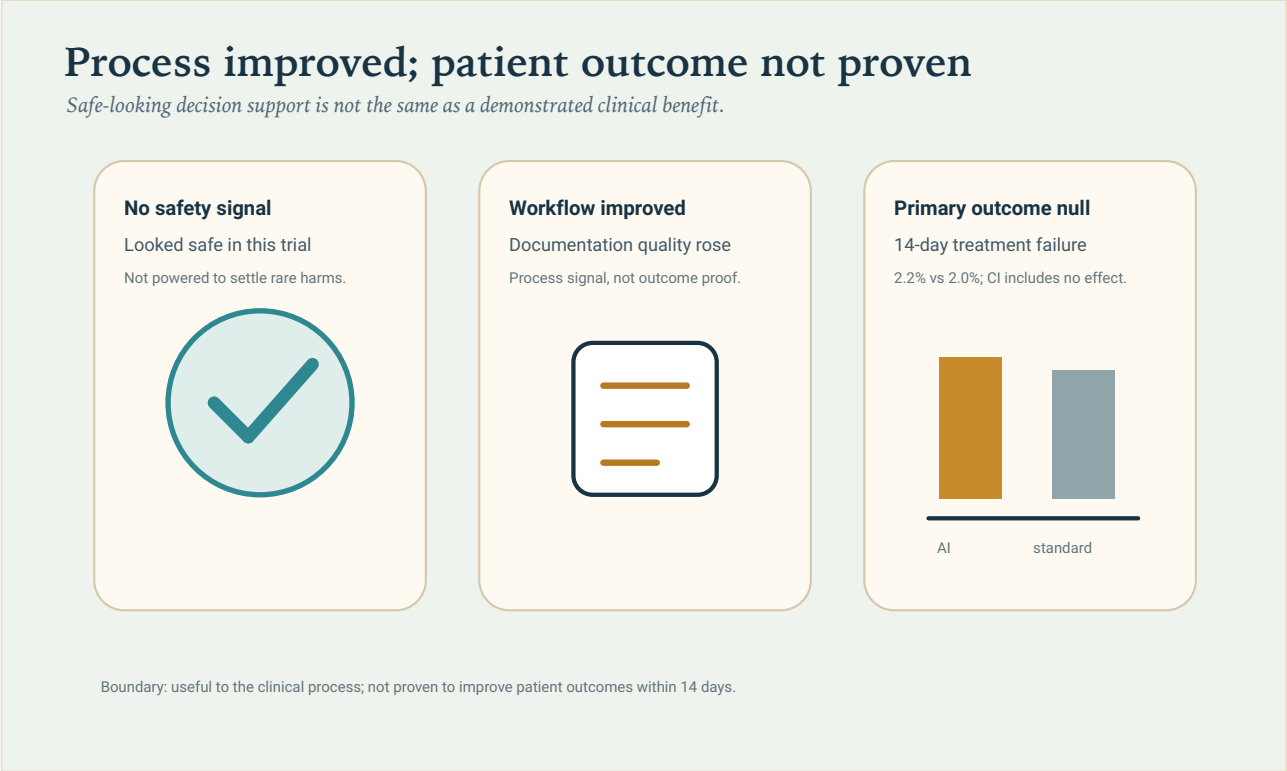
安全且有用，不等于有效

大多数关于医疗 AI 的标题都建立在错误类型的研究之上。一个模型在考试题目上得分很高，或者在精心策划的病例集上击败了医生，故事便不写自成：机器已经为临床做好了准备。

这项试验做了更困难也更罕见的事。它将一个生成式 AI 决策支持工具放入了真实的初级医疗中，在真实的设施中，面对真实的患者，并提出了那个真正重要的问题：患者是否获得了更好的结果?

诚实的答案是否定的——在 14 天内，没有可测量的改善。该工具没有显示出安全信号。它提高了临床文档的质量。它甚至可能降低了一些药物成本。但它并未显著减少治疗失败，作者谨慎地表示，任何益处如果存在，可能都是适度的。

这并不是研究的失败。这是研究在起作用。当医疗 AI 的证据是在患者身上而非基准测试上衡量时，负责任的证据就是这个样子的。



该工具未显示安全信号并改善了临床文档，但预先指定的 14 天患者结局并未显著改善。流程帮助与经证实的患者获益不是同一回事。

作者做了什么

团队在肯尼亚内罗毕和 Kiambu 县一家私人医疗网络（Penda Health）运营的 **16 家初级医疗设施**中进行了一项实用性、集群随机试验。这些设施中的医疗主要由**临床官员**——拥有三年文凭的中级从业

者——提供，通常无法方便地获得高级会诊。

随机化的单位是临床医生，而非患者。**103 名临床官员**被随机分配：52 名进入干预组，51 名进入对照组。两组使用同一套基于云的电子病历。干预组额外拥有「**AI Consult**」（2.0 版），一个基于 **OpenAI 的 GPT-4o** 大语言模型构建并嵌入该病历的决策支持工具。它读取临床医生记录的信息，并可标记诊断或治疗计划中可能存在的问题。临床医生保留完全自主权：他们可以接受、修改或忽略其建议。

> **是哪个模型，为什么这些细节很重要**
>
> 该工具是 **AI Consult 2.0**，运行 **OpenAI 的 GPT-4o**（2025 年 5 月发布），通过 OpenAI 的商业 API 在企业许可下访问，并以低随机性设置（温度 0.1）运行。它位于一个定制电子病历（EasyClinic 的 EMR）内部，并由旨在与肯尼亚国家治疗指南保持一致的系统提示词引导；作者发表了完整的指令提示词。
> 为什么要详细说明这些？因为结果是关于一个特定系统——一个模型版本、一个提示词、一个病历系统、一个环境——而非关于一般意义上的「医学中的大语言模型」。作者也提出了同样的观点：他们将其发现称为一个时间基准而非能力的固定估计。一个更新的模型、一个不同的提示词，或一个信息化程度较低的诊所，都可能改变结果。
> 关于独立性：OpenAI 后来提供了实物支持（云计算积分和关于使用其 API 的技术指导），但作者声明，使用 OpenAI 的决定是在该提议之前做出的，并且 OpenAI 在试验设计、数据收集、分析或发表决定中没有扮演任何角色。

在 2025 年 4 月 22 日至 7 月 16 日期间，**9,691 名患者**被纳入。**主要结局被刻意设计为以患者为中心的严格指标**：就诊后 14 天内由专家裁定的治疗失败复合指标——一个临床医生小组在不知晓研究分组的情况下，判断每位患者是否出现了不良结局，如未缓解或恶化的疾病。该试验已预先注册（泛非临床试验注册中心 202502499779176）。

这一设计选择正是关键所在。展示一个 AI 工具改变了临床医生写下的东西是容易的。展示它改变了患者身上发生的事情则困难得多，也重要得多。

他们发现了什么

主要结局并未改善。治疗失败发生在 AI 组 4,693 名患者中的 102 人（2.2%）和对照组 4,654 名患者中的 94 人（2.0%）。原始百分比在 AI 辅助下略微偏高，但调整后点估计偏向益处：**调整后的比值比（aOR）为 0.77（95% 置信区间 0.55 至 1.08，P = 0.13）**——无统计学显著性。原始数值与调整后数值之间的翻转并非计算错误；调整考虑了临床医生集群之间的差异。无论哪种情况，置信区间都舒适地包含「无效应」，因此无法声称任何益处，且以绝对数值来看，效应是微小的。

关于如何理解比值比、置信区间和 P 值的简明解释，请参阅临床结果阅读指南。

无安全信号——在限度之内。没有严重不良事件被判定与工具有关，独立审查未发现安全信号。作者对这一保证的上限是诚实的：该试验没有足够的统计功效来检测罕见的严重伤害，也没有预先指定的非劣效性或正式安全框架，因此它不能证明不常见事件的安全性。

文档质量得到了改善。在盲法专家审查的 2,000 次就诊中，使用 AI Consult 的临床医生在所有评分的

领域——记录的诊断、治疗计划和整体完整性——都产生了更好的临床文档。

处方几乎没有变化。在处方方面，包括正确的抗生素使用，没有显著差异（aOR 0.86，95% CI 0.48 至 1.55）。该工具并未改变抗生素处方率。

患者没有注意到差异。在完成满意度调查的 826 名患者中，两组之间的满意度基本相同，就诊时间也相似。

成本略微向下指向。在一项调整分析中，AI 组中抗生素相关成本更低——可能是通过更便宜的选择而非更少的处方——并且每位患者的抗生素节省似乎超过了运行该工具的每位患者成本。作者将此标记为提示性的，而非确定的：完整的总体拥有成本核算超出了试验范围。

作者自己的一行总结是最干净的版本：大语言模型辅助在这些限度内是安全的，但并未减少 14 天内的治疗失败，任何益处可能都是适度的。

为什么「无显著差异」不等于「它不起作用」

一个无效的主要结局很容易在任一方向上被过度解读。有两件事阻止了简单的故事。

首先，**该试验是为捕捉比其所发现的更大的效应而构建的**。初级医疗中严重的不良结局很罕见——此处约为 2%——因此将一个小型真实益处与噪声区分开需要极大的样本量。作者自己的事后统计功效计算表明，要检测他们观察到的那种大小的效应，需要一个**大得多的试验，大约需要超过 100,000 名患者**。在这种规模的研究中，一个非显著结果并不排除一个小型真实益处的存在；它意味着本研究无法解决这个问题。

其次，**对照组被部分模糊了**。一个配置错误曾短暂地让一些对照组临床医生也能访问 AI Consult，而共享网络中的临床医生会相互交流，并将习惯带过边界。这两种效应都倾向于使两组看起来更相似，将任何真实差异推向零。此外，宿主网络的运行标准已经相对较高，这给工具展示改善留下了更少空间。

这些都不是在援救「突破」标题。但它们确实意味着正确的解读是经过校准的，而非贬低的：在最困难和最诚实的终点上，该工具在两周内未明显帮助患者——但它确实在可测量的范围内帮助了病历记录，并且看起来是安全的。

这篇论文没有证明什么

- 它**没有**表明 AI 改善了患者结局。在主要 14 天终点上，没有显著益处。
- 它**没有**表明 AI 是无用的。点估计倾向于它，文档全面改善，药物成本趋势向下；无效结果与一个小型真实益处一致，只是试验规模太小无法确认。
- 它**没有**证明该工具对罕见伤害是安全的。它显示了无安全信号，但其统计功效或设计不足以认证不常见严重事件的安全性。
- 它**没有**表明「AI 击败医生」或取代临床医生。这是决策支持；临床医生保留了接受或拒绝的全部权力。
- 它**没有**自动泛化。该试验在肯尼亚单一城市私人网络中运行；农村、城郊和高收入环境可能在任

一方向上有所不同。
- 它**没有**确立成本节约。成本信号是提示性的，而非一项完整的经济学评估。

证据有多强?

对于核心主张——未证明短期患者伤害的减少——证据作为设计是强有力的，作为结论则是适当谦逊的。一项前瞻性、预先注册、集群随机化、采用盲法、以患者为单位的复合结局试验，接近你为这类工具能收集到的真实世界最佳证据。它比一个基准分数或一个病例集研究信息量大得多。

对于次要发现——更好的文档、不变的处方、相似的满意度、更低的抗生素成本——证据是良好的，但应作为次要发现来解读：支持性信号，而非标题，且易受相同污染和单一网络局限性的影响。

对于安全性，证据是令人安心的但有限度的：未发现信号，但这不是一项旨在发现罕见伤害的研究。

最有用的立场既不是「它有效」也不是「它失败了」。而是：一项谨慎的真实世界试验发现，这个 AI 工具未产生安全信号并改善了医疗流程，但未能在两周内展示患者结局的益处——而要检测任何此类益处将需要一项大得多的研究。

为什么重要

关于医疗 AI 的辩论正缺乏的正是这类证据。有成千上万的论文展示模型在考试中取得优异成绩并在整洁病例上与临床医生匹敌。但很少有大型、实用性、随机化试验衡量真实患者是否做得更好。这是其中之一，它落在了不光彩的真相上：通过考试与帮助患者并不是一回事。

这一差距就是整个故事。一个工具可以对临床医生真正有用——更清晰的笔记、更低的药费、第二双眼睛——但仍然无法在两周内推动一个硬性患者结局。这两个事实可以同时成立，而一个成熟的医疗系统必须将它们同时容纳，而非选择方便的那个。

它还将举证责任重置到一个有用的方向。如果一家公司想声称其临床 AI 改善了诊疗，相关的证据不是一个排行榜。而是一项像这样的试验，衡量对患者重要的结局——而且最好是一项更大的试验，因为这里的诚实教训是，适度的益处需要大规模的研究才能看到。

清洁摘要

一项在 16 家肯尼亚初级医疗设施中进行的实用性集群随机试验，测试了一个嵌入临床官员所用电子病历的生成式 AI 决策支持工具（「AI Consult」）。在 9,691 名患者中，专家裁定的 14 天内治疗失败复合指标在使用该工具时为 2.2%，未使用时为 2.0%（调整后比值比 0.77，95% CI 0.55–1.08，P = 0.13）——无显著差异。该工具未显示安全信号，改善了所有评分领域的临床文档质量，未改变处方，使患者满意度不变，并与略低的抗生素成本相关。作者得出结论，它在这些限度内是安全的，但并未减少治疗失败，任何益处可能都是适度的；要检测所观察到的那种大小的效应，将需要一项大得多的试验

（约 100,000 名患者规模）。该结果并未表明临床 AI 能改善患者结局，也未表明它是无用的——它表明，一个看起来安全并帮助了流程的工具，在两周内未能在单一城市私人网络中明显帮助患者。

No-BS 检查

论文所展示的：在一项真实世界随机试验中，将一个基于大语言模型的决策支持工具添加到初级医疗病历中未产生安全信号，改善了临床文档质量，未改变处方，且未显著减少一项严格的 14 天患者治疗失败复合指标。

合理但未经证实的：该工具产生了试验规模太小而无法检测的小型真实治疗失败减少；一旦计入全部成本它能节省资金；更好的文档最终能转化为更好的诊疗。

它没有展示的：临床 AI 能改善患者结局；它对罕见事件是安全的；它取代或超越了临床医生；这些结果可转移至农村或高收入环境；成本节约已经确立。

主要局限：统计功效针对比所观察到更大的效应量（罕见结局需要极大样本）；一个配置错误污染了对照组，共享网络中的习惯模糊了对比，两者都偏向于无差异；单一城市私人网络且标准已经较高；无预先指定的非劣效性或安全框架；仅限 14 天短期视野。

普通读者应该抱有多大信心？高度信心——该工具在本试验中未显示安全信号且改善了文档。高度信心——它在此处未明显改善 14 天患者结局。对于任何声称它作为一项患者获益干预「有效」或「失败」的说法，信心应低——这一问题确实尚未解决，需要一项大得多的研究。恰当的立场：一个关于一项工具的安全且改善了医疗流程的谨慎真实世界结果——而非一项判决 AI 能改变或毁掉初级医疗的裁决。

来源

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