The Clean Paper

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

Клинический AI, который выглядел безопасным и улучшил документацию — но не результаты пациентов

Прагматическое кластерно-рандомизированное испытание в 16 кенийских учреждениях первичной помощи проверило генеративный AI-инструмент поддержки решений, добавленный в медицинскую карту clinical officers. Среди 9 691 пациента строгий, подтверждённый экспертами 14-дневный комбинированный показатель неудачи лечения составил 2,2% с инструментом против 2,0% без него (скорректированное отношение шансов 0,77; 95% ДИ 0,55–1,08; P = 0,13) — значимой разницы нет. Инструмент не дал сигнала опасности, улучшил документацию по всем оценённым направлениям, не изменил назначения или удовлетворённость и показал тенденцию к меньшим затратам на антибиотики. Это не неудачное исследование, а редкое честное — измеренное на пациентах, а не бенчмарках — и оно приходит к неэффектной истине: инструмент, который выглядел безопасным и помогал процессу, не продемонстрировал пользы для пациентов за две недели, а для выявления возможной небольшой пользы потребуется гораздо более крупное испытание.

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

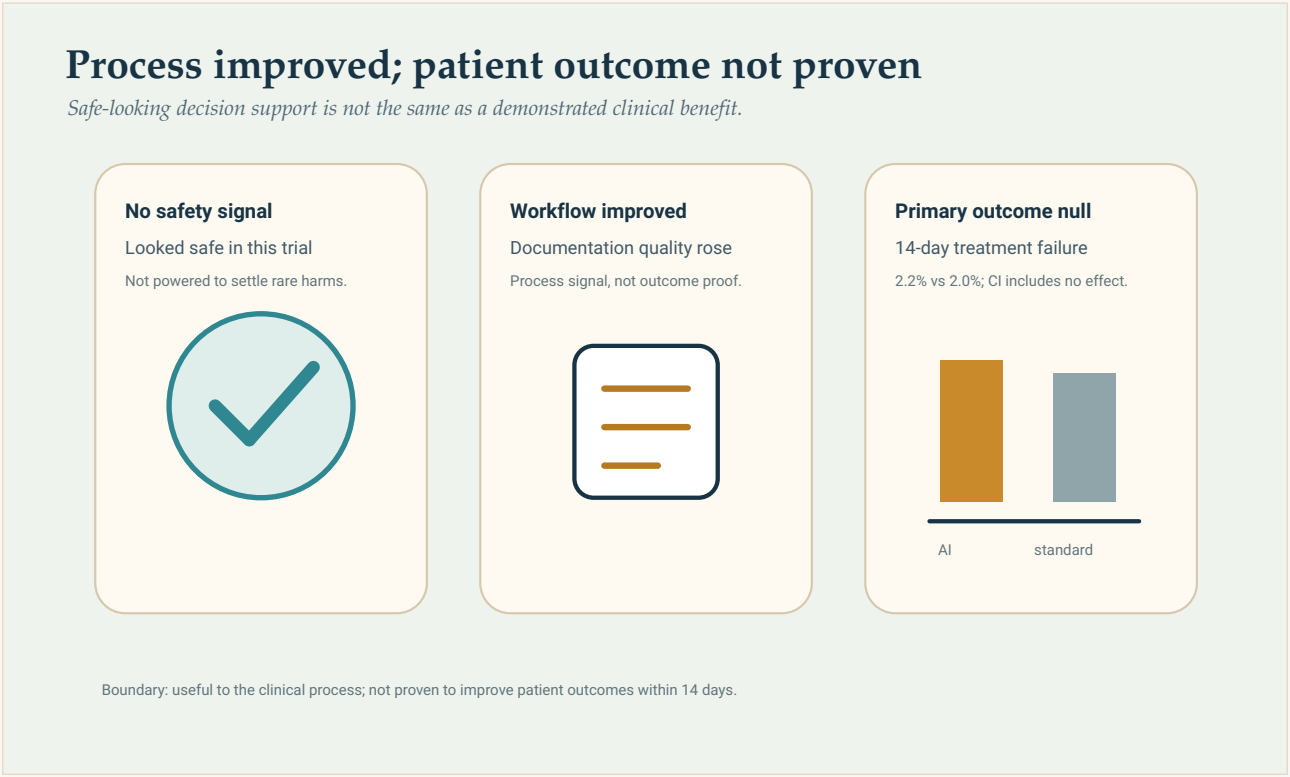
Безопасный и полезный — не то же самое, что эффективный

Большинство заголовков о медицинском AI вырастают из неправильного типа исследований. Модель хорошо отвечает на экзаменационные вопросы или превосходит врачей на специально подобранных клинических виньетках — и история пишется сама: машина готова к клинике.

Это испытание сделало более трудную и редкую вещь. Оно поместило инструмент поддержки решений на основе генеративного AI в настоящую первичную медицинскую помощь, в настоящие учреждения, с настоящими пациентами — и задало вопрос, который действительно имеет значение: пациентам стало лучше?

Честный ответ — нет. По крайней мере измеримо и в течение 14 дней. Инструмент не дал сигнала опасности. Он улучшил качество клинической документации. Возможно, даже немного снизил расходы на некоторые лекарства. Но значимого уменьшения неудач лечения не было, и авторы осторожно говорят: если польза существует, она, вероятно, скромна.

Это не провал исследования. Это исследование, которое работает как надо. Так выглядят ответственные доказательства об AI в медицине, когда измеряют последствия для пациентов, а не баллы бенчмарка.



Инструмент не дал сигнала опасности и улучшил клиническую документацию, но заранее определённый 14-дневный результат для пациентов значимо не улучшился. Помощь процессу — не то же самое, что доказанная польза для пациента.

Что сделали авторы

Команда провела прагматическое кластерно-рандомизированное испытание в **16 учреждениях первичной помощи** частной сети Penda Health в округах Найроби и Киамбу, Кения. Помощь там в значительной степени оказывают **клинические специалисты среднего звена** — медицинские работники среднего уровня с трёхлетним дипломом, часто без лёгкого доступа к консультации старших врачей.

Единицей рандомизации был клиницист, а не пациент. **103 клинические специалисты среднего звена** случайно распределили: 52 — в группу вмешательства, 51 — в контроль. Обе группы пользовались той же облачной электронной медицинской картой. Группа вмешательства дополнительно получила **AI Consult** (версия 2.0), инструмент поддержки решений на основе большой языковой модели **OpenAI GPT-4o**, интегрированный в эту карту. Он читал информацию, которую вносил клиницист, и мог указывать на возможные проблемы с диагнозом или планом лечения. Окончательное решение всегда оставалось за клиницистом: совет можно было принять, изменить или проигнорировать.

> **Какая именно модель — и почему детали важны**
>
> Инструментом был **AI Consult 2.0** на **OpenAI GPT-4o** (выпуск мая 2025 года), доступный через коммерческий API OpenAI по корпоративной лицензии и настроенный на низкую случайность (temperature 0,1). Он работал внутри специальной электронной медицинской системы EasyClinic EMR и управлялся системными промптами, написанными в соответствии с национальными клиническими рекомендациями Кении; авторы опубликовали полный инструктивный промпт.
> Зачем такая конкретика? Потому что результат относится к *одной конкретной системе* — одной версии модели, одному промпту, одной медицинской карте и одной среде, — а не к «LLM в медицине» вообще. Авторы подчёркивают то же самое: они называют свой результат *временным эталоном, а не фиксированной оценкой возможностей*. Более новая модель, другой промпт или менее цифровизированная клиника могли бы изменить результат.
> Что касается независимости: OpenAI позднее предоставила поддержку в натуральной форме (кредиты на облачные вычисления и технические советы по API), но авторы отмечают, что решение использовать OpenAI было принято *до* этого предложения и что OpenAI не участвовала в дизайне испытания, сборе или анализе данных либо решении о публикации.

Между 22 апреля и 16 июля 2025 года в исследование включили **9 691 пациента**. **Первичный результат намеренно был ориентирован на пациента и строг**: подтверждённый экспертами *комбинированный показатель неудачи лечения* в течение 14 дней после визита. Панель клиницистов, ослеплённая относительно группы, оценивала, был ли у пациента неблагоприятный результат — например, болезнь не прошла или

ухудшилась. Испытание заранее зарегистрировали (Pan-African Clinical Trials Registry 202502499779176).

Именно этот выбор дизайна — суть. Легко показать, что AI меняет то, что клиницист записывает. Гораздо труднее — и значительно важнее — показать, что он меняет то, что происходит с пациентом.

Что они обнаружили

Первичный результат не улучшился. Неудача лечения произошла у 102 из 4 693 пациентов (2,2%) в группе AI и у 94 из 4 654 (2,0%) в контроле. Сырые проценты были немного выше с AI, но после корректировки точечная оценка склонялась к пользе: **скорректированное отношение шансов 0,77 (95% доверительный интервал 0,55–1,08; P = 0,13)** — статистически незначимо. То, что направление меняется между сырыми и скорректированными числами, не арифметическая ошибка: корректировка учитывает различия между кластерами клиницистов. В любом случае доверительный интервал без проблем включает «никакого эффекта», поэтому заявлять пользу нельзя, а абсолютный эффект был крошечным.

Как читать отношение шансов, доверительные интервалы и P-значения вместе простым языком, смотрите в руководстве по чтению клинического результата.

Сигнала опасности не было — в пределах возможностей исследования. Ни одно серьёзное нежелательное явление не признали связанным с инструментом, а независимый обзор не нашёл сигнала безопасности. Авторы честно обозначают границу такого успокоения: испытание не имело достаточной статистической мощности для редких тяжёлых вредов и не имело заранее определённой схемы non-inferiority или формальной рамки безопасности, поэтому оно не может *доказать* безопасность относительно редких событий.

Документация стала лучше. Среди 2 000 консультаций, просмотренных ослеплёнными экспертами, клиницисты с AI Consult лучше документировали все оцениваемые компоненты — диагноз, план лечения и общую полноту записи.

Назначения почти не изменились. Значимой разницы в назначениях, включая правильное применение антибиотиков, не было (скорректированное отношение шансов 0,86; 95% ДИ 0,48–1,55). Частота назначения антибиотиков не изменилась.

Пациенты разницы не заметили. Среди 826 пациентов, заполнивших опрос удовлетворённости оценки были практически одинаковыми в обеих группах; длительность консультации тоже была похожей.

Расходы немного склонялись вниз. В скорректированном анализе расходы на антибиотики были ниже в группе AI — вероятно, из-за выбора более дешёвых препаратов, а не меньшего числа назначений, — и экономия на антибиотиках на одного пациента выглядела больше стоимости работы инструмента на одного

пациента. Авторы обозначают это как намёк, а не установленный факт: полный расчёт совокупной стоимости владения не входил в испытание.

Однострочное резюме самих авторов — самое чистое: помощь LLM в этих пределах выглядела безопасной, но не уменьшила неудачи лечения за 14 дней, а любая польза, вероятно, скромна.

Почему «нет значимой разницы» не означает «это не работает»

Нулевой первичный результат легко преувеличить в любую сторону. Две вещи не позволяют простой истории.

Во-первых, **испытание было спроектировано для обнаружения большего эффекта, чем тот, который оно увидело**. Серьёзные неблагоприятные исходы в первичной помощи редки — здесь около 2%, — поэтому для отделения небольшой реальной пользы от шума нужны огромные выборки. Постфактум расчёты мощности авторов показывают, что для обнаружения эффекта такого размера, как наблюдаемый, потребовалось бы **гораздо более крупное испытание — порядка более 100 000 пациентов**. Незначимый результат в исследовании такого масштаба не исключает небольшой реальной пользы; он означает, что это исследование не могло её различить.

Во-вторых, **сравнение частично размылось**. Из-за ошибки конфигурации некоторые клиницисты контрольной группы короткое время имели доступ к AI Consult, а сотрудники одной сети общаются и переносят привычки через границу между группами. Оба эффекта делают группы похожее и тянут любую настоящую разницу к нулю. Кроме того, сама сеть уже работала по относительно высоким стандартам, поэтому у инструмента было меньше пространства для улучшения.

Ничто из этого не спасает заголовок о «прорыве». Но правильное чтение должно быть калиброванным, а не пренебрежительным: по самому сложному и честному показателю этот инструмент не продемонстрировал пользы для пациентов за две недели — и одновременно заметно помог с документацией и выглядел безопасным.

Чего это *не* доказывает

- Это **не показывает**, что AI улучшил результаты пациентов. По первичному 14-дневному показателю значимой пользы не было.
- Это **не показывает**, что AI бесполезен. Точечная оценка была в его пользу, документация улучшилась, расходы на лекарства склонялись вниз; нулевой результат совместим с небольшой реальной пользой, для подтверждения которой исследование было

слишком маленьким.

- Это **не доказывает безопасность относительно редких вредов**. Сигнала опасности не нашли, но исследование не было спроектировано или достаточно мощно, чтобы сертифицировать безопасность для нечастых тяжёлых событий.
- Это **не показывает, что «AI побеждает врачей» или заменяет клиницистов**. Это *поддержка* решения; окончательная власть принять или отвергнуть совет была у клинициста.
- Результат **не обобщается автоматически**. Испытание проходило в одной частной городской сети Кении; в сельских, пригородных или более богатых системах результаты могут отличаться в любую сторону.
- Оно **не устанавливает экономию средств**. Сигнал затрат интересен, но это не завершённая экономическая оценка.

Насколько сильны доказательства?

Для центрального утверждения — что доказанного уменьшения краткосрочного вреда пациентам нет — дизайн доказательств силён, а вывод должным образом скромен. Проспективное, заранее зарегистрированное кластерно-рандомизированное испытание с ослеплённым комбинированным показателем на уровне пациента — почти лучший тип реальных доказательств, который можно получить для такого инструмента. Оно гораздо информативнее балла бенчмарка или исследования клинических виньеток.

Для вторичных результатов — лучшей документации, неизменных назначений, похожей удовлетворённости, меньших расходов на антибиотики — доказательства хороши, но их следует читать как вторичные: поддерживающие сигналы, а не главный вывод, и они подвержены тем же ограничениям из-за контаминации групп и одной сети.

Для безопасности доказательства успокаивают, но имеют чёткую границу: сигнала нет, однако это не было исследованием редкого вреда.

Самая полезная позиция — не «работает» и не «провалилось». Она такая: тщательное реальное испытание показало, что этот AI-инструмент не дал сигнала опасности и улучшил процесс оказания помощи, но не продемонстрировал пользы для пациентов в течение двух недель — и для обнаружения такой пользы, если она есть, потребуется значительно большее исследование.

Почему это важно

Дебату о медицинском AI отчаянно не хватает именно таких доказательств. Есть тысячи статей, где модели блестяще сдают экзамены или равняются клиницистам

на аккуратных задачах. Крупных прагматических рандомизированных испытаний, измеряющих, стало ли реальным пациентам лучше, — очень мало. Это одно из них, и оно приходит к неэффектной истине: сдать тест — не то же самое, что помочь пациенту.

Этот разрыв и есть вся история. Инструмент может быть действительно полезен клиницистам — лучшие записи, более дешёвые лекарства, ещё одна пара глаз — и одновременно не сдвинуть жёсткий клинический результат за две недели. Оба факта могут быть истинны одновременно, и зрелая система здравоохранения должна удерживать их вместе, а не выбирать удобный.

Работа также полезно сдвигает бремя доказательства. Если компания хочет заявить, что её клинический AI улучшает помощь, релевантное доказательство — не таблица лидеров. Это испытание вроде этого, с результатами, важными для пациентов, — и в идеале ещё большее, потому что честный урок здесь в том, что скромную пользу видно только в больших исследованиях.

Краткий итог

Прагматическое кластерно-рандомизированное испытание в 16 кенийских учреждениях первичной помощи проверило генеративный AI-инструмент поддержки решений AI Consult, добавленный в электронную карту клинические специалисты среднего звена. Среди 9 691 пациента подтверждённый экспертами комбинированный показатель неудачи лечения в течение 14 дней составил 2,2% с инструментом против 2,0% без него (скорректированное отношение шансов 0,77; 95% ДИ 0,55–1,08; P = 0,13) — значимой разницы нет. Инструмент не дал сигнала опасности, улучшил клиническую документацию по всем оценённым направлениям, не изменил назначения, не повлиял на удовлетворённость пациентов и ассоциировался с несколько меньшими расходами на антибиотики. Авторы заключают: в этих пределах он выглядел безопасным, но не уменьшил неудачи лечения; возможная польза, вероятно, скромна, и для эффекта наблюдаемого размера потребовалось бы гораздо более крупное испытание — порядка 100 000 пациентов. Результат не показывает ни того, что клинический AI улучшает результаты пациентов, ни того, что он бесполезен: он показывает, что инструмент, который выглядел безопасным и помогал процессу, не продемонстрировал пользы для пациентов за две недели в одной частной городской сети.

Проверка без преувеличений

Что показывает статья: В реальном рандомизированном испытании добавление LLM-инструмента поддержки решений к первичной помощи не дало сигнала опасности, улучшило качество клинической документации, не изменило назначения и не

уменьшило значимо строгий 14-дневный комбинированный показатель неудачи лечения.

Что правдоподобно, но не доказано: Что инструмент даёт небольшое реальное снижение неудач лечения, слишком малое для этого испытания; что он экономит деньги после учёта всех затрат; что лучшая документация в итоге превращается в лучшую помощь.

Чего работа не показывает: Что клинический AI улучшает результаты пациентов; что он безопасен относительно редких событий; что он заменяет или превосходит клиницистов; что результаты переносятся на сельские или более богатые системы; что экономия средств установлена.

Основные ограничения: Мощность рассчитывалась на больший эффект, чем наблюдаемый; редкие исходы требуют очень больших выборок. Ошибка конфигурации контаминировала контрольную группу, а общая сеть размывает различия — оба фактора смещают результат к нулю. Одна частная городская сеть с уже высокими стандартами; отсутствие заранее определённой non-inferiority или рамки безопасности; короткий 14-дневный горизонт.

Сколько доверия стоит иметь обычному читателю? Высокое в том, что в этом испытании инструмент не дал сигнала опасности и улучшил документацию. Высокое в том, что здесь он не продемонстрировал улучшения 14-дневных результатов пациентов. Низкое относительно любого общего вердикта «работает» или «не работает» как вмешательство для пользы пациентов — этот вопрос действительно не закрыт и требует значительно большего исследования. Уместная позиция: тщательный реальный результат об инструменте, который не дал сигнала опасности и улучшил процесс помощи, а не приговор, что AI преобразует — или разрушает — первичную медицину.

Источники

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