Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Retrospective testing and prospective randomized trial of guardrails and a verifier for AI care plans

Sanjay Basu

sanjay.basu@ucsf.edu

University of California San Francisco

Parth Sheth

University of Pennsylvania

Dan Wanek

Waymark

Bernardo Arevalo

Waymark

Aakriti Kinra

Waymark

Namrata Elemaran

Waymark

John Morgan

Virginia Commonwealth University

Rajaie Batniji

Waymark

Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11515604/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. The authors are employees of Waymark, a public benefit organization that provides free social and medical services for patients receiving Medicaid.

Abstract

Health systems are adding verifiers, AI models that check another AI's drafted care plans, to reduce errors and omissions. We tested a verifier in two ways. First, three AI models drafted care plans for 447 retrospective test cases; a verifier caught 5 of 17 errors checkable against the patient's record when counting flags for wrong recommendations, and 10 of 17 when also including flags for missing care. Second, in a prospective trial (NCT07718893), we randomized 240 patients in complex primary care management programs in four states to unassisted physicians, physicians with a guardrailed AI assistant, or physicians with the guardrailed assistant plus a verifier. The verifier raised 11 alerts across 77 encounters; physician adjudicators judged 9 correct, but clinicians in the trial acted on only 1. Safety events were similar across arms. Effective verification may therefore require alerts that clinicians must resolve, despite the risk of alert fatigue.

Introduction

Health systems now use AI models, specifically large language models, to draft clinical documents, including care plans. These drafts can contain errors that matter for patients. In the NOHARM benchmark, up to 24.6% of model consultations included a recommendation that could cause severe harm, and more than 80% of severe errors were omissions[1]. In adversarial tests, models expanded on made-up clinical details in 50% to 82% of cases[2]. Benchmarks such as MedHELM now measure these errors across 121 clinical tasks[3].

Health systems add two kinds of protection. The first is a set of guardrails around the drafting model: instructions that require cited sources and red flags, retrieval of guideline text, and automated checks of the draft. Groups that govern health AI, such as the Coalition for Health AI, ask that these tools be tested where they are used[4]. The second is a verifier, a second AI model that checks the first model's draft before it reaches the record. VeriFact checked statements in AI-written hospital summaries against the patient's record and agreed with clinicians about as often as clinicians agreed with each other[5]. SourceCheckup found that 50% to 90% of model answers to medical questions were not fully supported by the sources they cited[6]. In 39,849 primary care visits in Kenya, a tool that reviewed clinicians' notes was associated with 16% fewer diagnostic errors and 13% fewer treatment errors[7].

We did not find a large randomized trial in live care that tested either protection for AI-drafted care plans. Good benchmark scores do not ensure good results in care. In a pre-registered randomized study, members of the public who used AI models identified the correct condition no more often than controls, although the models alone scored well[8]; Agrawal and colleagues called this gap an evaluation illusion[9]. Prospective studies have tested diagnostic AI models in simulation[10,11], in a single-arm study of 98 primary care patients[12], and in randomized trials on written clinical cases[13,14]. Of 4,667 published medical AI studies, 113 (2.4%) were randomized[15], and 70 of 86 randomized trials of clinical AI (81%) reported a positive primary outcome[16].

To help in live care, a verifier must do three things. It must flag the errors that matter for the patient. It must stay quiet otherwise, because clinicians override 46.2% to 96.2% of decision-support alerts[17] and repeated alerts lower acceptance[18]. Its flags must also change what the clinician signs, which trust in a drafted plan, called automation bias, can prevent[19–21].

We built a verifier called CITE (Clinical Inference Tethered to Evidence). CITE is a second AI model that reads the drafted care plan before the physician signs it. It looks for wrong recommendations, statements that the patient's record does not support, and care that a guideline recommends but the draft leaves out. It shows each problem as a flag beside the draft, and the physician decides what to do. We tested CITE in two ways, with the same definitions: retrospectively, on AI-drafted plans for authored Medicaid test cases, and prospectively, in a pre-registered randomized trial in Medicaid primary care. The trial compared unassisted physicians, physicians with a guardrailed AI assistant, and physicians with the guardrailed assistant plus CITE. We asked two questions. First, how often do AI-drafted care plans contain errors checkable against the patient's record, and how many does CITE catch? Second, how do safety events, escalation, and 30-day acute care compare across the three arms, and how often do clinicians act on CITE's flags?

Results

Participants and test cases

The trial randomized 240 patients between 16 July and 17 September 2026: 79 to unassisted physicians (Arm 1), 82 to Gemini 3.1 Pro drafts with safety guardrails (Arm 2), and 79 to the same guardrailed drafts checked by CITE (Arm 3) (Fig. 1). Each patient had one encounter, and we analyzed all 240 by intention to treat. CITE ran in 77 of the 79 Arm 3 encounters. Patients had a mean age of 42.8 years (SD 16.1); 15 were younger than 18 and 8 were 65 or older. They came from Ohio (117), Virginia (61), Washington (41), and Illinois (21) (Table 1). We linked 237 of 240 patients (98.8%) to 30-day acute care data: 79, 81, and 77 by arm. The retrospective test included 1,341 care plans: three AI models each wrote one plan for each of 447 authored Medicaid test cases.

Table 1
Baseline characteristics by arm

Characteristic	Arm 1: unassisted (n = 79)	Arm 2: guardrails (n = 82)	Arm 3: guardrails + CITE (n = 79)	Total (n = 240)
Age, years, mean (SD) (n = 237)	40.5 (15.9)	44.5 (15.6)	43.4 (16.6)	42.8 (16.1)
Younger than 18 years, n	5	4	6	15
65 years or older, n	2	3	3	8
Female, n (n = 237)	62	57	57	176
Male, n	17	24	20	61
Sex unknown, n	0	1	2	3
Black or African American, n	31	38	31	100
White, n	17	10	14	41
Race unknown or not reported, n	28	28	31	87
State: Ohio / Virginia / Washington / Illinois, n	36 / 21 / 14 / 8	44 / 17 / 15 / 6	37 / 23 / 12 / 7	117 / 61 / 41 / 21

Age and sex come from Waymark records for the 237 patients linked to 30-day data. Other race categories (Hispanic or Latinx 9, American Indian or Alaska Native 2, Other 1) are not shown by arm. State comes from the patient’s Medicaid plan.

Checkable errors in test cases and live care

We call an error that a simple rule could check against the patient’s recorded diagnoses, medications, allergies, and laboratory values a checkable error. We based its five types on errors that rule-based checks already target: drug–disease, drug–allergy, and dose checks in order-entry decision support[22], explicit criteria for inappropriate and omitted therapy[23], and omission of regular medications[24] (Methods). The trial registration calls these errors codifiable hazards. Table 2 follows each study from the plans that CITE checked to the actions that clinicians took.

Table 2
From plans checked to clinician action, in test cases and live care

Step	Authored test cases (retrospective)	Live care (trial, Arm 3)
Plans checked by CITE	1,341 (3 models × 447 cases)	77 Gemini 3.1 Pro drafts
Plans with a checkable error	17/1,341 (1.3%; 95% CI 0.7 to 2.0)	0/77 (the 1 live error, among 161 drafts, was in Arm 2)
Errors caught, primary threshold (wrong recommendations)	5/17 (29.4%; 95% CI 10.3 to 56.0)	Not estimable (no errors)
Errors caught, broader threshold (adds missing care)	10/17 (58.8%; 95% CI 32.9 to 81.6)	Not estimable (no errors)
Specificity, primary threshold	1,288/1,324 (97.3%; 95% CI 96.3 to 98.1)	70/70 (100.0%; 95% CI 94.9 to 100.0)
Specificity, broader threshold	1,243/1,324 (93.9%; 95% CI 92.5 to 95.1)	65/70 (92.9%; 95% CI 84.1 to 97.6)
Plans with any CITE flag	193/1,341 (14.4%)	7/77 (9.1%)
CITE flags per plan, mean	0.19	0.14
Flags judged valid by physician adjudicators	Not adjudicated	9/11
Valid flags acted on in the signed note	Not applicable	1/9

Both studies used the same error definitions, verifier code, and model. A checkable error is one of five types of patient-specific error that a rule could check against the patient's record (Methods). Live-care specificity uses the pre-specified denominator, which excludes 7 drafts without enough clinical content; Supplementary Table 8 gives values for all 77 drafts. The 9 valid flags in live care were missed workups and unsupported statements, which fall outside the five types.

In the test cases, 17 of 1,341 plans (1.3%; 95% CI 0.7 to 2.0) had a checkable error: 16 of 447 Gemini 3.1 Pro plans (3.6%), 1 of 447 GPT-5.6 plans (0.2%), and 0 of 447 Claude Opus 4.8 plans (Fig. 2a). In live care, blinded physicians found a checkable error in 1 of 161 Gemini 3.1 Pro drafts (0.6%; 95% CI 0.0 to 3.4). That draft was in Arm 2, so none of the 77 drafts that CITE checked had a checkable error.

CITE accuracy, alert burden, and clinician action

CITE raises three types of flags: wrong recommendations, missing care, and unsupported statements. The pre-registered primary threshold counts flags for wrong recommendations. The broader threshold also counts flags for missing care. Neither threshold counts flags for unsupported statements.

In the test cases, CITE caught 5 of 17 checkable errors (29.4%; 95% CI 10.3 to 56.0) at the primary threshold, with a specificity of 1,288/1,324 (97.3%; 95% CI 96.3 to 98.1). At the broader threshold, CITE

caught 10 of 17 (58.8%; 95% CI 32.9 to 81.6), with a specificity of 1,243/1,324 (93.9%; 95% CI 92.5 to 95.1) (Table 2).

In live care, sensitivity could not be estimated, because no draft that CITE checked had a checkable error. CITE raised no flags for wrong recommendations, so it had no false alarms at the primary threshold. Specificity was 70/70 (100.0%; 95% CI 94.9 to 100.0) in the pre-specified denominator, which excluded 7 drafts without enough clinical content to assess. At the broader threshold, specificity was 65/70 (92.9%; 95% CI 84.1 to 97.6).

In all, CITE raised 11 flags in 7 of 77 encounters (9.1%): median 0, mean 0.14, and maximum 4 flags per encounter (Fig. 2b), within the pre-registered ceiling of a median of 3. Physician adjudicators judged 9 of 11 flags clinically valid (81.8%; 95% CI 48.2 to 97.7; AC1 0.846): all 5 flags for unsupported statements and 4 of 6 flags for missing care, from 6 encounters (Table 3). None of the 9 valid flags was a checkable error. The valid flags for missing care were missed workups: evaluation of a suspected first seizure in a patient taking an anticoagulant, a bedside neurological examination for suspected cauda equina syndrome, colonoscopy for two weeks of rectal bleeding, and a blood count for rectal bleeding in Crohn's disease. The valid flags for unsupported statements included an interpreter need and a heart failure exacerbation that the record did not support (Supplementary Table 6).

Table 3
The 11 CITE flags in live care, by type

Flag type	Flags, n	Judged valid by physician adjudicators, n	Valid flags acted on in the signed note, n
Missing care (indicated action left out)	6	4	1 (action added in part)
Unsupported statement (not supported by the record)	5	5	0 (flagged statement retained in all 5)
Total	11	9	1

CITE showed its flags in real time beside the draft; the trial did not record whether clinicians opened each flag. The signed final notes acted on 1 of the 9 valid flags (Fig. 2c): the note added colonoscopy, in part, for 1 of 4 valid flags for missing care. All 5 statements that CITE flagged as unsupported, from 2 encounters, remained in the signed notes.

Escalation, safety events, and time to decision

Safety events were similar across arms. The safety composite, which counts a missed do-not-miss diagnosis, under-triage, a contraindicated medication, or a missed escalation, occurred in 4 of 79 encounters (5.1%) in Arm 1, 4 of 82 (4.9%) in Arm 2, and 5 of 79 (6.3%) in Arm 3 (Fig. 3). The Arm 2 minus Arm 1 difference was − 0.2 points (95% CI − 8.0 to 7.5), and the Arm 3 minus Arm 2 difference was 1.5 percentage points (95% CI − 6.4 to 9.7). Physician agreement on the composite was substantial (Fleiss κ 0.728; 95% CI 0.555 to 0.859; components in Supplementary Table 4).

Over-escalation, the protocol's safety outcome, occurred in 0 of 79 Arm 1 encounters, 1 of 82 Arm 2 encounters, and 0 of 79 Arm 3 encounters. The Arm 3 minus Arm 2 difference was − 1.2 percentage points (95% CI − 6.6 to 3.5); the upper bound was below the pre-specified 5-point margin, and the midpoint review did not trigger the halt rule. Appropriate escalation and median time to decision were similar across arms (Supplementary Table 8).

Unresolved errors and referrals within 30 days

A checkable error remained in 2 signed final plans, 1 in Arm 2 and 1 in Arm 3, and neither was corrected within 30 days (Supplementary Table 8). A referral in the final plan was acted on within 30 days for 7 of 29 (24.1%) in Arm 1, 12 of 50 (24.0%) in Arm 2, and 13 of 45 (28.9%) in Arm 3 (Arm 3 minus Arm 2, 4.9 percentage points; 95% CI − 16.8 to 26.5; Supplementary Note 2).

30-day acute care and adverse events

Within 30 days, an emergency department visit occurred in 16 of 79 linked Arm 1 patients (20.3%), 14 of 81 Arm 2 patients (17.3%), and 19 of 77 Arm 3 patients (24.7%). Hospitalization occurred in 9 of 79 (11.4%), 6 of 81 (7.4%), and 4 of 77 (5.2%), respectively. The Arm 2 minus Arm 1 differences were − 3.0 percentage points for emergency visits (95% CI − 15.1 to 9.2) and − 4.0 points for hospitalization (95% CI − 13.7 to 5.5). The Arm 3 minus Arm 2 differences were 7.4 points for emergency visits (95% CI − 5.3 to 20.0) and − 2.2 points for hospitalization (95% CI − 10.7 to 6.2). No deaths were recorded within 30 days. Sensitivity analyses did not change these results (Supplementary Table 3).

Discussion

We tested a verifier for AI-drafted care plans in two ways and report two findings. First, in plans for authored test cases, CITE caught 5 of 17 checkable errors at its primary threshold and 10 of 17 at its broader threshold, with specificities of 97.3% and 93.9%. In the trial, it raised no false alarms at its primary threshold, and physicians judged 9 of its 11 flags valid. Second, safety events were similar across arms, and over-escalation stayed within the pre-specified margin, but clinicians acted on only 1 of the 9 valid flags.

We found that the trial was underpowered for its main question. Checkable errors were much rarer in live care than in the test cases: 1 of 161 live drafts (0.6%) had one, compared with 16 of 447 test-case plans (3.6%) from the same model. That one error was in an Arm 2 draft, which CITE did not check, so the trial could not estimate how many live errors CITE catches. CITE's 9 valid flags were missed workups and unsupported statements, which are real problems outside the five checkable types. We also found a key challenge for use in real-world care. Even when CITE found real problems, clinicians seldom carried its flags into the signed care plan. Future research needs to examine why flags are or are not incorporated into care plans, including the design of the tool's user interface and the balance between alert fatigue and automation bias.

The comparison of the guardrailed assistant with unassisted physicians answers the first question that health systems ask: can a guardrailed AI model draft plans for physician review without adding safety events? We did not find a larger randomized comparison of this kind in live care; the closest prospective study was a single-arm study of 98 patients[12]. Guardrailed drafts were similar to unassisted physicians on safety events (4.9% vs 5.1%), escalation, and time to decision, and emergency visits and hospitalizations were numerically lower with guardrails. This is the local evidence that governance groups ask for[4]. The confidence intervals do not exclude differences of several points in either direction.

Two explanations could account for the gap between valid flags and signed notes. All 5 flagged unsupported statements stayed in the signed notes, and 1 of 4 valid flags for missing care led to a change. One explanation is automation bias[19,20]. In a randomized trial of 44 physicians trained in AI literacy, wrong AI suggestions lowered diagnostic accuracy by 14.0 percentage points (95% CI − 18.9 to − 9.1)[21]. In the randomized substudy of NOHARM, physicians with AI help often left out useful AI recommendations[1], and other randomized studies of AI decision support found similar gaps[13,25]. Another explanation, which the trial could not test, is that clinicians judged some flagged items not to need a change. Two designs could close this gap: a flag that the clinician must resolve before signing, and a flag sent to the care team member who acts on it, such as the pharmacist for a medication problem.

The lower error rate in live drafts held even in rising-risk patients with many conditions and medications. The drafting model, chosen for its benchmark accuracy, also had the most clinically consequential errors of the three models in the test cases (15.0% of plans). The test cases over-represented risky scenarios and used a different rating process, so the two rates are not directly comparable. The direction agrees with reports that benchmark results do not carry over directly to care[9,26,27].

CITE raised few flags, and this matters for any design that makes clinicians resolve them. Clinicians override most decision-support alerts[17], acceptance falls with repeated alerts[18], and a trial of alerts for acute kidney injury found harm at some hospitals[28]. CITE flagged 9.1% of live encounters and 14.4% of plans for authored test cases, did not increase over-escalation, and did not detectably lengthen time to decision.

This study has limitations. Beyond the low error rate in live drafts, the trial had 80% power only for an increase in the composite of about 13.7 points (post hoc). The emergency visit rate was 7.4 points higher with CITE than with guardrails alone (95% CI − 5.3 to 20.0), and the hospitalization rate was 2.2 points lower; both are exploratory results with wide intervals, as in other small randomized trials of clinical AI[13,29]. The trial did not record whether clinicians viewed, accepted, or overrode each flag, so we measured action from the signed note. In the retrospective test, Claude Opus 4.8 drafted the ratings and was also the verifier's model and one of the rated models, which may favor its own plans and its own detections. One model drafted all live care plans, in one organization in four states.

In plans for authored test cases, CITE caught 5 of 17 checkable errors at its primary threshold and 10 of 17 at its broader threshold. In the trial, CITE raised 11 flags across 77 encounters, physicians judged 9 valid, and clinicians acted on 1; safety events were similar across arms. Effective verification may therefore require flags that clinicians must resolve, designed to limit alert fatigue, and trials that select encounters with more checkable steps, such as visits for anticoagulation or chronic kidney disease.

Methods

Study design and registration

We report two linked studies that used the same hazard definitions and the same verifier: retrospective testing of CITE on authored Medicaid test cases, and prospective validation in a pre-registered three-arm randomized trial. The trial was registered at ClinicalTrials.gov (NCT07718893; submitted 13 July 2026, before the first enrollment on 16 July 2026), and the analysis plan was posted at the Open Science Framework on 13 July 2026 (https://osf.io/vb5f3). The registered co-primary outcomes were CITE's diagnostic accuracy against blinded clinician adjudication, and clinician alert burden and acceptance. The protocol also specified over-escalation with a 5-percentage-point safety margin and a midpoint halt rule, appropriate escalation, and time to decision. We report the trial according to CONSORT 2025[30] and CONSORT-AI[31], and the retrospective testing with reference to TRIPOD-LLM[32]. Patients and the public were not involved in the design, conduct, or reporting of the trial.

Trial setting and participants

The trial took place in Waymark's multidisciplinary primary care service for Medicaid managed-care members in Ohio, Virginia, Washington, and Illinois. Waymark teams include community health workers, care coordinators, pharmacists, and therapists, supervised by physicians. Waymark enrolls members whom validated risk models identify as rising risk for emergency department visits and hospitalization[33,34]. Encounters came from the care team's clinical rounds, which by default discuss patients with multimorbidity or polypharmacy; this enriched enrollment toward these patients, as pre-specified. Eligible patients were Medicaid-enrolled and attributed to a participating primary care site, with English-language documentation and an encounter that required clinical reasoning. We excluded patients with hospice- or palliative-care-exclusive care plans, patients enrolled in a competing AI-safety study within the prior 90 days, administrative-only or pharmacy-only encounters, and encounters in which the supervising clinician was the principal investigator. The sample size of 240 was fixed in the approved protocol. Specificity and alert burden can be estimated when checkable errors are absent; sensitivity needs errors, so the trial could estimate it only if enough errors occurred.

Randomization and trial arms

We randomized patients 1:1:1 with a deterministic HMAC-SHA256 permuted-block scheme (block size 6), stratified by site. Allocation was concealed in the trial software. Physicians knew the arm from the drafting interface and were not blinded; outcome adjudicators were blinded.

In Arm 1, the supervising physician wrote the assessment and plan without AI assistance. In Arm 2, Gemini 3.1 Pro drafted the assessment and plan under safety guardrails, and the physician reviewed, edited, and signed it. The guardrails had six parts (Supplementary Table 7): a safety system prompt that required a cited source and a confidence level for every recommendation and prominent red flags and safety hazards; retrieval of matching summaries from a curated library of society guidelines; a citation check that marked citations outside that library as unverified; four automated verification checks (drug interactions, laboratory interpretation, guideline compliance, and a diagnostic safety net); a second-opinion consensus check; and a privacy filter that replaced patient identifiers. The model ran at temperature 0.2. These guardrails represent the safety layer that health systems commonly add on top of a language model. In Arm 3, Gemini 3.1 Pro drafted the plan under the same guardrails, and CITE then checked the draft as a further layer before physician review. We chose Gemini 3.1 Pro as the drafting model because it had the highest MedQA accuracy (97.4%) among frontier models in a comparison of general-purpose models and clinical AI tools[35]. In all arms the supervising physician signed the final note when closing the chart.

The CITE verifier

CITE (version cite-1.0) sends the language-model draft and structured patient context (age, sex, problems, medications, laboratory values, and history) to a second language model (Claude Opus 4.8). A fixed verification prompt asks the model to find three types of problems: wrong recommendations, statements that the record does not support, and missing care, meaning actions that a high- or moderate-certainty guideline recommendation requires for the patient but the draft omits. CITE does not retrieve guideline text at run time; the model applies its own guideline knowledge and labels each flag for missing care with the certainty of the evidence. Each flag carries a quote from the draft, a one-line rationale, and a severity; flags for missing care take their severity from the certainty label, with high certainty set to 3, and other flags have severity 2. A flag of severity 3 sets the displayed status to ESCALATE. The displayed status depends on severity, whereas the pre-registered thresholds depend on flag type. CITE showed its status and flags in real time in a side panel beside the draft, and the physician kept full decision authority. CITE returned results with a median latency of 1.7 s across the 73 of 77 runs with a stored timing.

Error definitions and outcomes

A checkable error, registered as a codifiable hazard, is a patient-specific error in a care plan that a rule could check against the patient's own documented diagnoses, medications, allergies, and laboratory values. We counted five types, based on errors that rule-based checks already target[22–24]: a recommended drug that the patient's diagnosis or laboratory value contraindicates; a drug that conflicts with a documented allergy; a guideline-indicated therapy omitted for an active diagnosis without a documented reason; a documented high-risk medication, such as an anticoagulant, insulin, an antiepileptic, or an antiretroviral, dropped without a documented reason; and a stated dose, threshold, or quantity that current guidance refutes. Each error was rated on a 4-point scale (1, minor; 2, would change

management; 3, plausible moderate harm; 4, plausible serious harm), and an error counted at severity 2 or higher. A checkpoint is one patient-specific opportunity for a checkable error, such as an active diagnosis with an indicated therapy, a recommended drug, or a documented allergy. In the retrospective testing, a plan was deficient if it contained at least one clinically consequential error of any kind (an omission, an inappropriate action, or an unsupported statement) at severity 2 or higher; checkable errors are the subset of deficiencies in the five types (Supplementary Note 1).

Co-primary outcome 1 was CITE's sensitivity, specificity, and predictive values for language-model drafts that contained at least one checkable error, against the majority judgment of three blinded physicians; as pre-specified, drafts without enough clinical content to assess were excluded from this denominator. The pre-registered primary (ESCALATE-only) threshold counts flags for wrong recommendations: a commission (a wrong or contraindicated recommendation), an allergy conflict, or a refuted numeric claim. The pre-registered broader (ANY-flag) threshold also counts flags for missing care (omissions). Co-primary outcome 2 was alert burden, measured as flags per encounter against a ceiling of a median of 3, and clinician acceptance of flags, measured as action in the signed note: a flag for missing care was acted on when the note added the flagged action, and a flag for an unsupported statement was acted on when the note removed the statement. Secondary outcomes were clinician action on CITE flags by flag type, unresolved checkable errors per checkpoint in the final plan, correction of these errors within 30 days, and completion of referrals within 30 days. Exploratory outcomes were the four-component safety composite (missed do-not-miss diagnosis, under-triage, contraindicated medication, or missed escalation) and 30-day emergency department visits and hospitalizations (Supplementary Table 1).

Authored test cases

We authored 447 whole-person Medicaid test cases from de-identified care activities, sampled across eight domains, including contraindication, omission, longitudinal follow-up, social needs, and true-negative and hard-negative cases, and frozen before generation (Supplementary Table 5). Three models (Gemini 3.1 Pro, GPT-5.6, and Claude Opus 4.8) each wrote one care plan per case under the same clinical-assistant prompt, for 1,341 plans. A language model drafted a rating for each plan under a pre-specified rubric with the definitions above, blinded to the generating model, with separate physician and community health worker role prompts, and three physicians reviewed and approved each rating. CITE checked each plan with the same verifier code and model as in the trial, and we computed its accuracy at the same two pre-registered thresholds.

Adjudication

Three board-certified physicians, blinded to arm, scored the safety composite, over-escalation, and appropriate escalation for every encounter from the signed note. The majority of three defined each outcome separately, so an encounter could count as a missed escalation in the composite without a majority judging that escalation was warranted. For the registered outcomes, three physician adjudicators scored redacted packets independently: the 161 language-model drafts for checkable

errors (the reference standard for outcome 1), the 240 final plans for unresolved checkable errors, and the 11 CITE flags for validity and presence in the final note. The draft and final-plan packets did not show arm, model, or verifier output. The flag packet revealed Arm 3 membership, so adjudicators scored it last. Before full scoring, the adjudicators scored the same 10 drafts and met to agree on definitions.

For the 30-day secondary outcomes, one physician reviewed Waymark records, including notes, tasks, admission, discharge, and transfer events, and claims, for the 30 days after each of the 78 encounters in which adjudicators listed a candidate referral or found an unresolved checkable error in the final plan; the chart reviewer confirmed referrals in 73 of them. The extracts did not show arm. A second physician independently reviewed 16 of these encounters. A referral was acted on if the record showed a completed visit, a visit with a different service for the same need, or a scheduled appointment; referrals to Waymark pharmacists, therapists, and community health workers counted the same as outside referrals.

30-day outcomes

We linked patients to Waymark's admission, discharge, and transfer feed by patient identifier, or by keyed hashes of the identifier or of name and date of birth when the identifier was not captured. An emergency department visit was an episode with patient class E, and a hospitalization was an episode with patient class I, within the 30 days after the index encounter. Deaths came from the feed and from payer eligibility records.

Statistical analysis

We report risk differences with Newcombe 95% CIs[36], Wilson 95% CIs for proportions of encounters, and exact Clopper–Pearson 95% CIs for diagnostic accuracy, flag validity, error prevalence, and referrals. We report agreement as Fleiss κ and Gwet AC1[37], because κ is low when nearly all items fall in one category[38]. We compared time to decision between Arm 3 and Arm 2 with the Mann–Whitney test, because 64 encounters had a recorded time of 0 s, and with the protocol's t-test on log seconds among encounters with a recorded time above 0 s. The differences in unresolved checkable errors per checkpoint and in referrals acted on used an encounter-level bootstrap (2,000 resamples), because one encounter can contribute several checkpoints or referrals. We did not adjust exploratory outcomes for multiplicity. The 30-day analyses used the 237 linked patients, and we did not impute missing outcomes.

Declarations

Ethics

WCG IRB approved the trial under protocol 20261776.

Use of large language models

Gemini 3.1 Pro and Claude Opus 4.8 were part of the interventions under study. In preparing this manuscript, the authors used Claude (Anthropic) to draft text, write analysis code, and produce tables and figures. The authors reviewed and edited all content and take responsibility for it. No language model chose the study design, the analysis specification, or the interpretation of the results, and no language model is an author.

Competing Interests

The authors are employees of Waymark, a public benefit organization that provides free social and medical services for patients receiving Medicaid.

Funding:

None.

Author Contribution

S.B. conceived the study and wrote the initial draft of the manuscript. P.S., D.W., B.A., A.K., N.E., J.M., and R.B. revised the manuscript. All authors approved the final manuscript.

Acknowledgements

None.

Data Availability

Individual participant data, including de-identified data, are not available. They derive from Medicaid records, which are subject to federal and state privacy restrictions. Aggregate results are posted at ClinicalTrials.gov (NCT07718893).

Code availability

The analysis code and the deployed CITE verifier, including its verification prompt, are available at https://github.com/sanjaybasu/cite-trial.

References

1. Wu, D. et al. First, do NOHARM: a medical safety benchmark and randomized study of physician and AI teaming on clinical consultations. Preprint at https://arxiv.org/abs/2512.01241 (2026).
2. Omar, M. et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun. Med.* 5, 330 (2025).
3. Bedi, S. et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nat. Med.* 32, 943–951 (2026).
4. Shah, N. H. et al. A nationwide network of health AI assurance laboratories. *JAMA* 331, 245–249 (2024).
5. Chung, P. et al. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI* 3, AIdbp2500418 (2026).
6. Wu, K. et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nat. Commun.* 16, 3615 (2025).
7. Korom, R. et al. AI-based clinical decision support for primary care: a real-world study. Preprint at https://arxiv.org/abs/2507.16947 (2025).
8. Bean, A. M. et al. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nat. Med.* 32, 609–615 (2026).
9. Agrawal, M. et al. The evaluation illusion of large language models in medicine. *npj Digit. Med.* 8, 600 (2025).
10. Tu, T. et al. Towards conversational diagnostic artificial intelligence. *Nature* 642, 442–450 (2025).
11. McDuff, D. et al. Towards accurate differential diagnosis with large language models. *Nature* 642, 451–457 (2025).
12. Brodeur, P. G. et al. Conversational diagnostic artificial intelligence in ambulatory primary care: a prospective feasibility study. Lancet https://doi.org/10.1016/S0140-6736(26)01535-7 (2026).
13. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* 7, e2440969 (2024).
14. Goh, E. et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat. Med.* 31, 1233–1238 (2025).
15. Wang, Q. et al. A quantitative analysis of global AI medical studies: gaps in randomized controlled trials. *npj Digit. Med.* 9, 403 (2026).
16. Han, R. et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit. Health* 6, e367–e373 (2024).
17. Poly, T. N. et al. Appropriateness of overridden alerts in computerized physician order entry: systematic review. *JMIR Med. Inform.* 8, e15653 (2020).
18. Ancker, J. S. et al. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med. Inform. Decis. Mak.* 17, 36 (2017).
19. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* 19, 121–127 (2012).

20. Lyell, D. & Coiera, E. Automation bias and verification complexity: a systematic review. *J. Am. Med. Inform. Assoc.* 24, 423–431 (2017).
21. Qazi, I. A. et al. Automation bias in large language model–assisted diagnostic reasoning among physicians trained in AI literacy: a randomized clinical trial. *NEJM AI* 3, AIoa2501001 (2026).
22. Kuperman, G. J. et al. Medication-related clinical decision support in computerized provider order entry systems: a review. *J. Am. Med. Inform. Assoc.* 14, 29–40 (2007).
23. O'Mahony, D. et al. STOPP/START criteria for potentially inappropriate prescribing in older people: version 2. *Age Ageing* 44, 213–218 (2015).
24. Cornish, P. L. et al. Unintended medication discrepancies at the time of hospital admission. *Arch. Intern. Med.* 165, 424–429 (2005).
25. Spitzer, P. et al. The effect of medical explanations from large language models on diagnostic decisions in radiology. *npj Digit. Med.* 9, 333 (2026).
26. van de Sande, D. et al. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. *npj Digit. Med.* 7, 58 (2024).
27. Marwaha, J. S. & Kvedar, J. C. Crossing the chasm from model performance to clinical impact: the need to improve implementation and evaluation of AI. *npj Digit. Med.* 5, 25 (2022).
28. Wilson, F. P. et al. Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. *BMJ* 372, m4786 (2021).
29. Osmanodja, B. et al. Randomized trial of electronic health record implemented AI risk prediction in kidney transplantation. *npj Digit. Med.* 9, 373 (2026).
30. Hopewell, S. et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. *BMJ* 389, e081123 (2025).
31. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* 26, 1364–1374 (2020).
32. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* 31, 60–69 (2025).
33. Patel, S. Y., Baum, A. & Basu, S. Prediction of non emergent acute care utilization and cost among patients receiving Medicaid. *Sci. Rep.* 14, 824 (2024).
34. Basu, S. et al. Clinical decision support for population health management: development and validation of integrated acuity and intervention prediction models. *J. Am. Med. Inform. Assoc.* 33, 670–678 (2026).
35. Vishwanath, K. et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nat. Med.* 32, 2405–2409 (2026).
36. Newcombe, R. G. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat. Med.* 17, 873–890 (1998).
37. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* 61, 29–48 (2008).

38. Feinstein, A. R. & Cicchetti, D. V. High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* 43, 543–549 (1990).

Figures

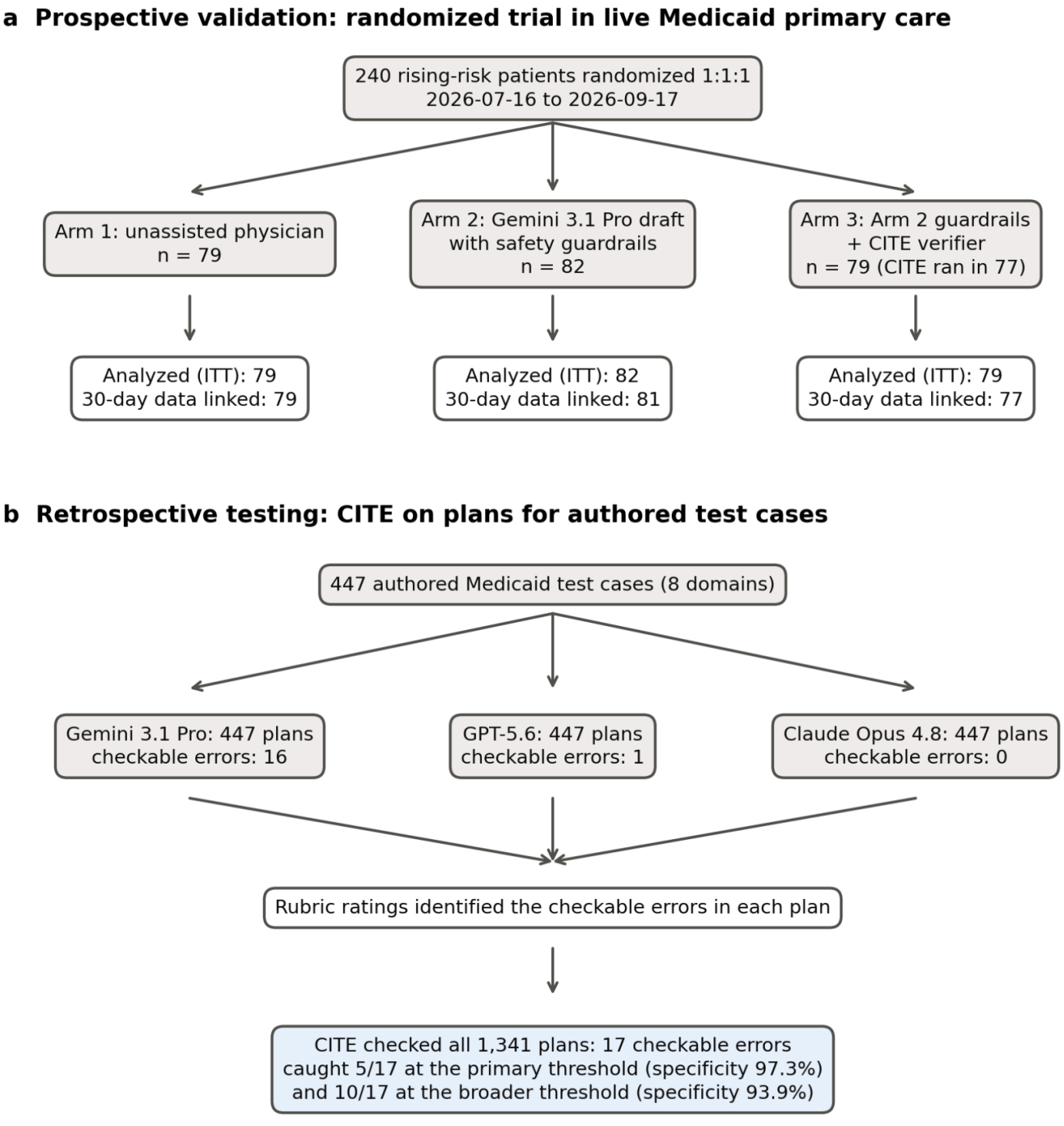


Figure 1

Study design and participant flow. a, Prospective validation in the randomized trial (NCT07718893): 240 rising-risk patients randomized 1:1:1 to unassisted physicians (Arm 1, n = 79), Gemini 3.1 Pro with safety guardrails (Arm 2, n = 82), or the same model plus the CITE verifier (Arm 3, n = 79; CITE ran in 77). All patients were analyzed by intention to treat; 30-day acute care data were linked for 79, 81, and 77 by arm. **b**, Retrospective testing on 1,341 language-model care plans for 447 authored Medicaid test cases. Three models each wrote 447 plans. Rubric ratings, drafted by a language model and approved by three physicians, identified 17 checkable errors (16, 1, and 0 by model). A checkable error is a patient-specific error that a rule could check against the patient's documented diagnoses, medications, allergies, and laboratory values (Methods). CITE checked every plan with the same verifier as in the trial and caught 5 of 17 errors at its primary threshold and 10 of 17 at its broader threshold.

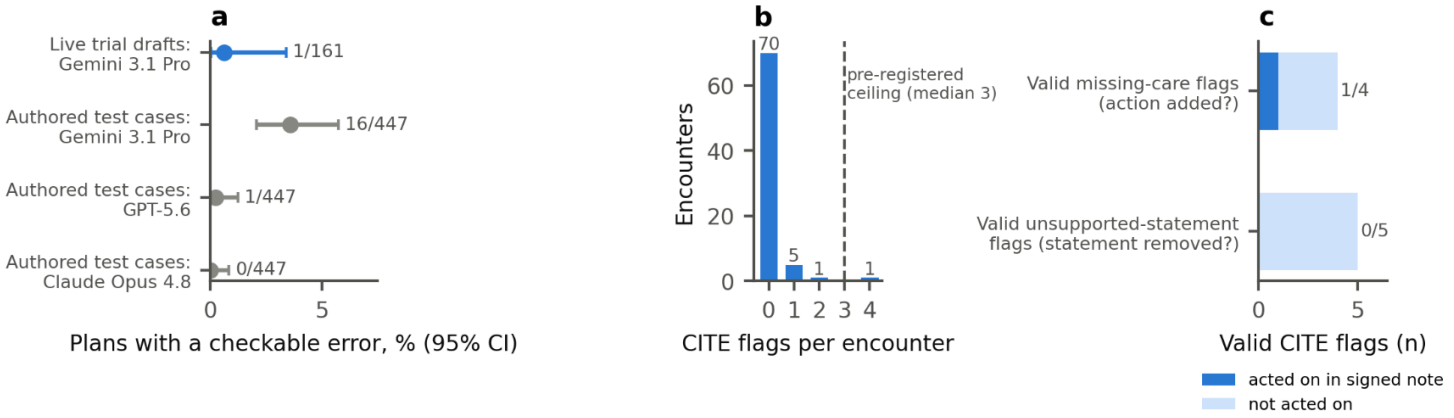


Figure 2

Error rates, flags, and clinician action. a, Proportion of care plans with a checkable error, with exact 95% CIs: live trial drafts from Gemini 3.1 Pro (1/161), and authored test cases by model (16/447 Gemini 3.1 Pro, 1/447 GPT-5.6, 0/447 Claude Opus 4.8). The two studies differ in case mix and rating method. **b**, CITE flags per encounter in the 77 Arm 3 encounters where CITE ran (11 flags; median 0); the dashed line marks the pre-registered ceiling of a median of 3. **c**, The 9 flags that physician adjudicators judged valid, by type, and whether the signed final note acted on them (an added action for flags for missing care; a removed statement for flags for unsupported statements).

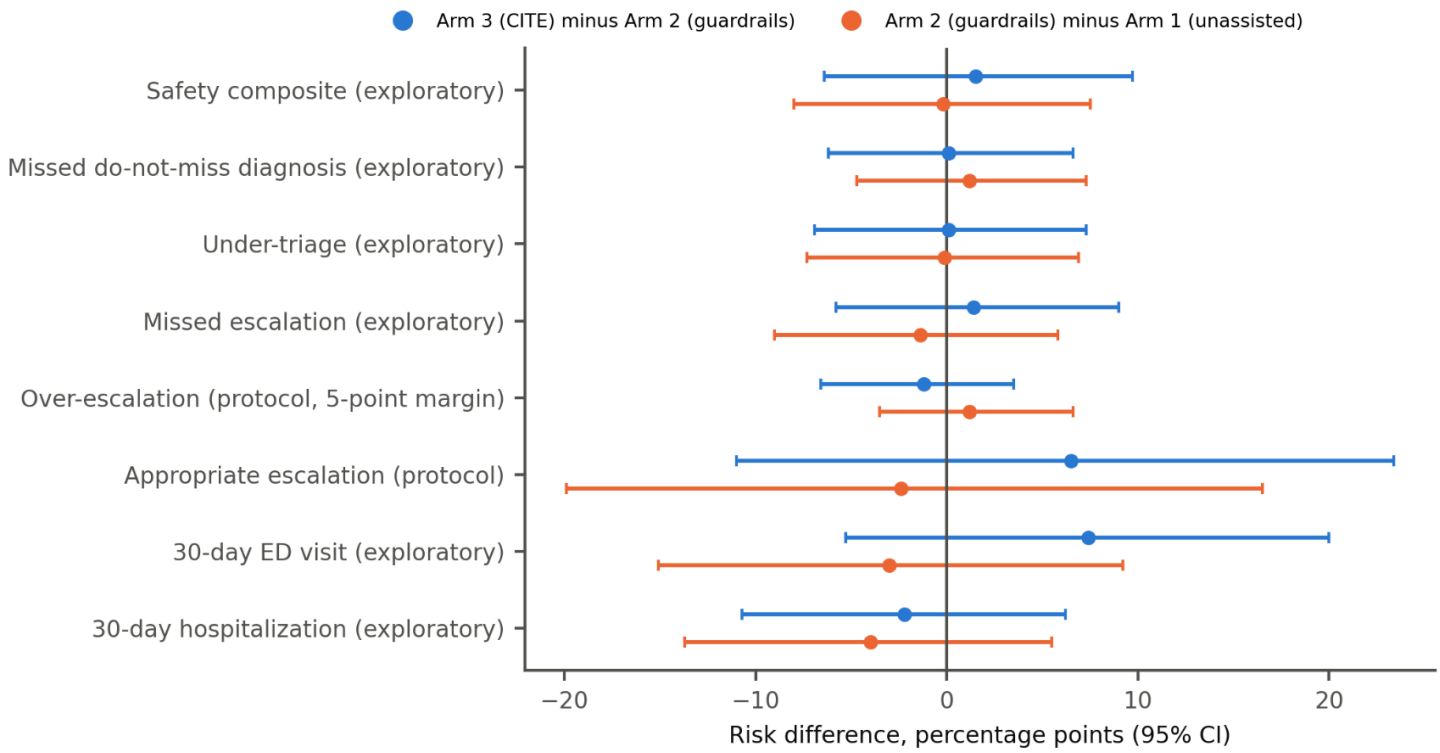


Figure 3

Escalation, safety, and acute care outcomes. Risk differences in percentage points with Newcombe 95% CIs for Arm 3 (CITE) minus Arm 2 (guardrails) and Arm 2 minus Arm 1 (unassisted).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BasuCITEnpjHSSupplementaryInformation.docx