

Prospective evaluation of a large language model clinical decision support system in the emergency department

Received: 1 March 2026

Accepted: 22 July 2026

Published online: 19 August 2026

 Check for updates

Liron Leibovitch ^{1,7}, Adi Ahituv ^{1,2,7}, Alon Gorenshtein ³, Dvir Aran ^{4,5},
Moran Sorka^{2,4}, Keren Miron¹ & Shahar Shelly ^{1,2,6} 

Prospective evidence for artificial intelligence (AI)-based clinical decision support in emergency departments remains limited. Here we conducted a DECIDE-AI stage 1 evaluation of SHAKED, a clinical decision support system built on multiple large language models, in a tertiary emergency department. Over 4 weeks, 1,138 patients were analyzed across two parallel units—one using SHAKED and one following routine rotations. Clinical adoption of SHAKED declined from 68% to 30%, owing to workload-sensitive disengagement (OR = 0.72 per shift hour, 95% CI 0.62 to 0.83). Physicians preferred the use of SHAKED for radiology consultations (OR = 2.98, 95% CI 1.58 to 5.63). No adverse events were detected, and expert review rated 99 of 100 sampled outputs as clinically appropriate. Emergency department length of stay did not differ between wings (4.9 h in both, $P = 0.99$). Intention-to-treat analysis showed a non-significant trend toward shorter consultation cycle time (−9.4 min, $P = 0.077$). These findings suggest that sustained clinician engagement, rather than algorithmic accuracy, may be the key barrier to effective clinical AI use in emergency departments. They inform randomized trial design but do not justify clinical deployment of AI clinical decision support at this stage. ClinicalTrials.gov identifier: [NCT06902675](https://clinicaltrials.gov/ct2/show/study/NCT06902675).

Emergency departments (EDs) face increasing patient volumes and prolonged throughput times, requiring clinicians to make time-sensitive decisions while integrating fragmented information from multiple sources^{1,2}. Delays often arise not from knowledge gaps but from coordination challenges in clinical workflows, in particular around specialty consultation initiation, prioritization and completion^{3,4}.

Most artificial intelligence (AI) applications in emergency medicine have focused on prediction tasks such as risk stratification or outcome forecasting^{5,6}. These systems are typically evaluated in retrospective datasets or silent deployments and are only loosely integrated into real clinical workflows⁷. As a result, prospective evidence showing

measured operational benefits such as reduced consultation time or diagnostic accuracy, and critically, evidence of how such systems perform during actual implementation, remains scarce^{8–10}.

The development pathway for AI-based clinical decision support systems (CDSS) requires staged evaluation that accounts for adoption dynamics, workflow integration and safety concerns^{11,12}. The DECIDE-AI framework¹³ defines early-stage clinical evaluation (stage 1) as small-scale, live assessment focusing on four objectives: (1) proof of clinical utility at small scale, (2) safety profile characterization, (3) human factor evaluation and (4) preparation for larger summative trials. This stage precedes definitive efficacy trials and serves to identify

¹Department of Neurology, Rambam Health Care Campus, Haifa, Israel. ²Faculty of Medicine, Technion Israel Institute of Technology, Haifa, Israel.

³Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA. ⁴Faculty of Biology, Technion – Israel Institute of Technology, Haifa, Israel. ⁵Taub Faculty of Computer Science Technion – Israel Institute of Technology, Haifa, Israel. ⁶Department of Neurology, Mayo Clinic, Rochester, MN, USA. ⁷These authors contributed equally: Liron Leibovitch, Adi Ahituv. ✉e-mail: shahar.shelly@technion.ac.il

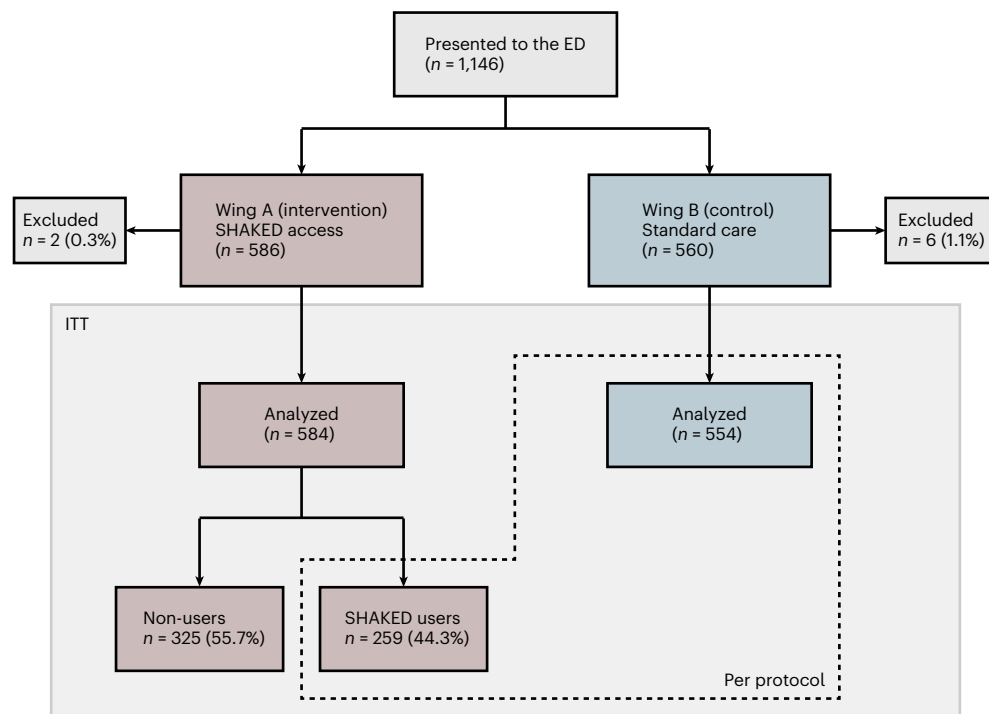


Fig. 1 | Study design and patient flow. CONSORT-style flow diagram showing study enrollment, strictly alternating wing allocation and patient classification. Of 1,146 enrolled patients, 586 were allocated to wing A (intervention) and 560 to wing B (control). Eight patients without specialty consultations were

excluded (two from wing A, six from wing B), yielding 584 and 554 for analysis. In wing A, 259 patients (44.3%) had physicians who opened the SHAKED interface (per-protocol population). The four-level exposure classification is detailed in Extended Data Fig. 3.

implementation barriers, refine the intervention, characterize adoption dynamics and generate preliminary effect estimates to inform sample size calculations for subsequent randomized controlled trials.

Unlike traditional predictive models that output risk scores from structured data, large language models (LLMs) can process unstructured clinical notes and retrieve condition-specific guidelines^{14,15}, potentially supporting documentation, communication and decision-making tasks that require narrative synthesis^{16,17}. Whether LLM-based systems improve outcomes through non-specific effects such as heightened awareness or workflow modification, or require active physician engagement with AI-generated recommendations, is unknown^{18,19}. This distinction has direct implications for training requirements, workflow integration strategies and scalability.

We conducted a prospective controlled pilot study to evaluate SHAKED during initial clinical implementation in a live tertiary-care ED. Following DECIDE-AI stage 1 objectives, we aimed to (1) assess implementation feasibility through adoption rate measurement and characterization of usage patterns, (2) identify human factors and implementation barriers that determine real-world engagement, (3) characterize the safety profile under live clinical conditions and (4) estimate preliminary causal effect using instrumental variable analysis to inform the design of a definitive randomized controlled trial. We hypothesized that active engagement would be associated with reduced consultation cycle time through a human-in-the-loop mechanism, but that real-world adoption would be partial and influenced by workload, clinical context and temporal factors.

Results

Study population and allocation balance

Between 9 November and 7 December 2025, we analyzed 1,146 consecutive adult patients presenting to the ED (Fig. 1). Patients were assigned by strictly alternating allocation at triage to wing A, in which treating physicians had access to the SHAKED CDSS, or to wing B,

Table 1 | Baseline characteristics by wing assignment (allocation balance)

Variable	Wing A (n=584)	Wing B (n=554)	SMD	P
Age, years, mean±s.d.	63.9±18.5	64.5±19.4	0.027	0.372
Female sex, n (%)	284 (48.6)	260 (46.9)	0.034	0.607
P1 (immediate), n (%)	15 (2.6)	6 (1.1)	0.111	0.101
P2 (emergent), n (%)	4 (0.7)	5 (0.9)	0.025	0.937
P3 (urgent), n (%)	448 (76.7)	412 (74.4)	0.055	0.395
P4 (less urgent), n (%)	109 (18.7)	125 (22.6)	0.096	0.120
P5 (non-urgent), n (%)	7 (1.2)	2 (0.4)	0.10	0.179
Radiology consult, n (%)	421 (72.1)	408 (73.6)	0.035	0.601
Hospital admission, n (%)	265 (45.4)	258 (46.6)	0.024	0.731
Diagnosis count, mean±s.d.	5.1±4.9	4.9±4.8	0.035	0.310

All SMDs <0.12. P values are from two-sided Mann–Whitney U tests (age and diagnosis count) and chi-square tests (categorical variables; Fisher's exact test for priority 5), comparing wing A with wing B, unadjusted for multiple comparisons.

which received routine care without SHAKED access. Eight patients without specialty consultations were excluded (<1%), leaving 584 in wing A and 554 in wing B. Mean age was 64.2 ± 18.9 years; 47.8% were female. Triage acuity distribution was 1.8% immediate (priority 1), 0.8% emergent (priority 2), 75.6% urgent (priority 3), 20.5%

Medical History Summary
Censored

Censored patient information

Medical History Summary: [Copy Summary] [Highlighter] [Copy Highlighted Text] [Clear All]

Censored presents with **dyspnea and productive cough** since yesterday [P1].

Relevant Medical History

Underlying oncologic disease: The patient has a known history of **pulmonary adenocarcinoma with brain metastases**, diagnosed **Censored** [H38]. He underwent **whole-brain radiation therapy (WBRT)** to two parietal metastases (right and left) receiving 2700 cGy in 3 fractions [H38]. He subsequently received **chemotherapy**, with the last cycle administered **Censored** [H37]. A PET-CT from November 2025 showed a known small LLL mass with residual uptake consistent with active disease, radiation-related changes, improved hilar lymphadenopathy, and residual pathologic uptake in a 6 mm lymph node and a mediastinal node up to 10 mm short axis. Brain MRI from November 2025 showed **marked improvement in metastases** [H75]. The patient is followed at an oncology clinic (Zevulun) [H75].

Recurrent respiratory decompensation: The patient has a documented baseline **room-air oxygen saturation of approximately 77-87%**, requiring **home oxygen supplementation** (nasal cannula), established during a hospitalization in July-August 2023 [H52].

Most recent hospitalization **Censored** The patient was admitted approximately 6 months ago with **bronchopneumonia** (organism unspecified) and **acute kidney injury** (creatinine peaked at 1.54, baseline ~1.17) [H75]. He presented with generalized weakness, chills, and worsening cough and dyspnea. On admission: BP 113/67, HR 97, temperature 38.5 degrees C, room-air saturation 87% [H75]. Chest auscultation revealed bilateral crackles, more prominent on the right. Respiratory viral panel (influenza A/B, RSV, COVID-19, parainfluenza, adenovirus, metapneumovirus, rhinovirus) was **negative** [H71]. Blood cultures were **negative** [H69, H70]. CBC on **Censored** showed WBC 12.88 x10³/L (neutrophils 87.5%, absolute neutrophils 11.28), HGB 12.8 g/dL, platelets 91 x10³/L [H68]. He was treated with IV ceftriaxone followed by oral levofloxacin 750 mg daily (complete **Censored** [H75]. Renal function improved spontaneously; creatinine at discharge was 1.27 [H73]. He was discharged home in good condition on intermittent nasal cannula oxygen [H76].

Prior hospitalization **Censored** Admitted with **Bordetella pertussis** (IgA positive, IgG negative), presenting with fever and productive cough of 3 weeks' duration [H52]. Room-air saturation was up to 86%; home oxygen was arranged during this admission [H52]. CBC on admission showed severe pancytopenia: WBC 1.96, HGB 8.2 g/dL, platelets 67, absolute neutrophils 0.90 x10³/L [H34] — attributed to chemotherapy effect. Respiratory cultures showed mixed upper respiratory flora; no bacterial pathogen isolated [H45]. He was managed conservatively (antibiotic course with roxithromycin had already been completed in the community) and discharged with droplet isolation instructions [H52].

COPD: Diagnosed in 2010; CT chest (December 2024) showed bronchiectatic changes in bilateral basal lobes and small pleural nodules [H75]. Spirometry from 2017 showed FVC 65% predicted, FEV1 58% predicted, FEV1/FVC 91% [H25], suggesting a restrictive or mixed pattern. He received inhaled ipratropium and salbutamol during the July 2025 hospitalization [H50, H51].

Current presentation **Censored** Vital signs show HR 110 (above normal range of 60-100), BP 15/72 mmHg (systolic value appears erroneous as documented), temperature 36.7 degrees C, **room-air saturation 82%**, saturation on supplemental oxygen 94%, pain score 0, full consciousness [P2]. Laboratory studies show elevated fasting glucose 147 mg/dL (normal 70-99) [P8], serum creatinine 1.10 (normal 0.7-1.3) [P8], BUN 17.0 (normal 6.4-25.7) [P8], sodium 137 (normal 133-145) [P8], chloride 103.0 (normal 98-110) [P8], potassium was cancelled [P8]. Complete blood count shows WBC 9.93 (normal 4-10.8) [P7], hemoglobin 13.2 g/dL (normal 13-17) [P7], platelets 170 (normal 130-500) [P7], with elevated neutrophils 82.0% (normal 40-71) [P7] and low lymphocytes 6.1% (normal 18-44) [P7]; RDW 16.03 (normal 11.5-15) [P7]. Arterial blood gas shows pH 7.35 (borderline low, normal 7.35-7.45) [P5], pCO2 80 mmHg (elevated, normal 35-45) [P5], pO2 15 mmHg (severely low, normal 75-90) [P5], HCO3 33.1 (elevated, normal 22-26) [P5], base excess -5.60 (elevated, normal -2 to 2) [P5], lactate 1.40 (slightly elevated, normal 0-1.3) [P5], ionized calcium 1.13 (low, normal 1.15-1.35) [P5], glucose 149 mg/dL (elevated, normal 70-110) [P5]. The triage nurse referred the patient to internal medicine [P3].

Differential Diagnoses [Regenerate] [Delete version]

Community-acquired pneumonia (bacterial) / Acute bronchopneumonia

Highest likelihood AND high harm-if-missed. This patient has a near-identical presentation to his January 2026 admission: productive cough with acute onset (since yesterday), tachycardia (HR 110), and significant desaturation (SpO2 82% on room air, 94% on supplemental O2). His baseline room-air SpO2 is 77-87%, so the current 82% represents a meaningful acute-on-chronic decline. He has multiple risk factors for pneumonia: metastatic pulmonary adenocarcinoma, COPD with bronchiectatic changes, prior chemotherapy (last cycle July 2025), active smoking (~30 pack-years), and recurrent pneumonia history. Temperature is currently 36.7 (afebrile), but he was afebrile on some prior admissions as well. The January 2026 episode was treated successfully with ceftriaxone then levofloxacin. New labs [P8] confirm normal renal function (creatinine 1.10, BUN 17) and normal electrolytes, supporting safe antibiotic dosing without renal dose adjustment. CBC [P6/P7] shows neutrophilia (NEUT% 82%, WBC 9.93) with lymphopenia (LYMPH% 6.1%) — a pattern consistent with bacterial infection and/or stress response. Elevated RDW (16.03) may reflect underlying nutritional or chronic disease effect. Blood gas [P4/P5] shows type 2 respiratory failure (pO2 15, pCO2 60, pH 7.35, HCO3 33) with metabolic compensation, consistent with acute-on-chronic respiratory decompensation. Estimated likelihood: 55-65%. Harm-if-missed: HIGH.

History: H26, H75, H52, H68, P1, P2, P6, P7, P8 **Literature:** 2,38495, 2,38497, 2,38498, 2,38499, 2,34201, 2,38533

Acute-on-chronic COPD exacerbation (infectious or non-infectious trigger)

Second most likely. The patient has documented COPD since 2010 with bronchiectatic changes on CT, and a prior history of exacerbations. Productive cough and worsening dyspnea are classic AECOPD features. Tachycardia (HR 110) and desaturation (SpO2 82% room air) are consistent with moderate-to-severe exacerbation. Blood gas [P4/P5] confirms type 2 respiratory failure: pCO2 60 mmHg (elevated, consistent with CO2 retention), pH 7.35 (low-normal, compensated), HCO3 33.1 (elevated, indicating chronic metabolic compensation), pO2 15 (severely low). This ABG pattern is highly consistent with COPD-related hypercapnic respiratory failure. The distinction from pneumonia may be difficult clinically and both may coexist. Neutrophilia (NEUT% 82%) and lymphopenia (LYMPH% 6.1%) on CBC [P6] support an infectious trigger. Estimated likelihood: 20-30%. Harm-if-missed: HIGH.

History: H52, H75, H26, P1, P2, P4, P5, P6 **Literature:** 2,38338, 2,38340, 2,38343, 2,38344, 2,38345

Pulmonary embolism

Must not be missed despite lower probability. This patient has metastatic malignancy (a major hypercoagulable risk factor), tachycardia (HR 110), and acute hypoxemia (SpO2 82% room air) with acute-onset dyspnea. He is also relatively immobile given his oncologic disease burden. The absence of pleuritic chest pain and the presence of productive cough make PE less likely than pneumonia, but PE can present atypically in cancer patients. Blood gas [P4/P5] shows severely low pO2 (15 mmHg) which, while consistent with COPD/pneumonia, is also compatible with PE-related V/Q mismatch. Normal renal function [P8] (creatinine 1.10) supports safe contrast administration for CTPA if PE probability rises. Estimated likelihood: 10-15%. Harm-if-missed: CATASTROPHIC.

History: H26, P1, P2, P4, P8 **Literature:** 2,38697, 2,34211, 2,38343, 2,38700, 2,38351

Fig. 2 | Representative SHAKED system output. Representative SHAKED output as viewed by the treating physician. The screen shows the autogenerated medical history summary, differential diagnoses and suggested workup for one patient encounter.

less urgent (priority 4) and 0.8% non-urgent (priority 5)²⁰. Hospital admission occurred in 46.0% of encounters, with balance between wings (45.4% wing A versus 46.6% wing B, standardized mean difference (SMD) = 0.02). Radiology consultations accounted for 72.9% of all consultations (Table 1). Seventy-one wing B patients (12.7%) were treated by physicians who also practiced in wing A during the study period (without SHAKED access during wing B encounters); these were retained in the control group per intention-to-treat (ITT) principles. System logs confirmed that no physician treating wing B patients accessed the SHAKED interface during those encounters. Input data completeness was high for structured fields (>99%), but no SHAKED processing was triggered for 52 wing A encounters (8.9%), classified as the ‘no-exposure’ group.

Eight physicians treated patients in wing A; all completed pre-deployment training, and none had prior experience with AI-based clinical decision support. The same physicians also treated patients in wing B as part of routine rotation. Individual physician identifiers were not recorded owing to institutional data governance requirements. Baseline covariate balance was confirmed (all SMDs <0.12, allocation AUC = 0.543), with null falsification tests and no temporal drift in controls (Supplementary Table 6 and Extended Data Fig. 1). Wing A non-users ($n = 325$) showed no faster baseline consultation times

compared with wing B controls (-6.1 min, $P = 0.337$), confirming that SHAKED users were not preselected for inherently faster workflows.

Adoption, usage patterns and human factors

Clinicians engaged SHAKED (representative system output; Fig. 2) in 44.3% of eligible encounters (259/584 patients in wing A; 95% confidence interval (CI) 40.3% to 48.4%). Adoption rates declined over the 4-week study period: study week 1, 67.9% (108/159); study week 2, 39.4% (54/137); study week 3, 37.9% (53/140); and study week 4, 29.7% (44/148) (Fig. 3a). Adoption declined by approximately 7% per study day (odds ratio (OR) = 0.93 per day, 95% CI 0.91 to 0.95, $P < 0.001$). Technical system failures occurred in 44 encounters (17% of the 259 encounters in which physicians opened SHAKED; 8% of all 584 wing A patients).

Adoption showed substantial workload sensitivity. Each additional hour of ED arrival time reduced engagement likelihood by 28% (OR = 0.72 per hour, 95% CI 0.62 to 0.83, $P < 0.001$), adjusting for patient acuity and consultation type (Supplementary Table 3). Physicians used the system preferentially for radiology consultations (OR = 2.98, 95% CI 1.58 to 5.63, $P < 0.001$) compared to other consultation types, despite no measured differential effectiveness across consultation types (interaction $P = 0.35$). Owing to the unavailability of physician identifiers, it was impossible to assess individual physician learning curves.

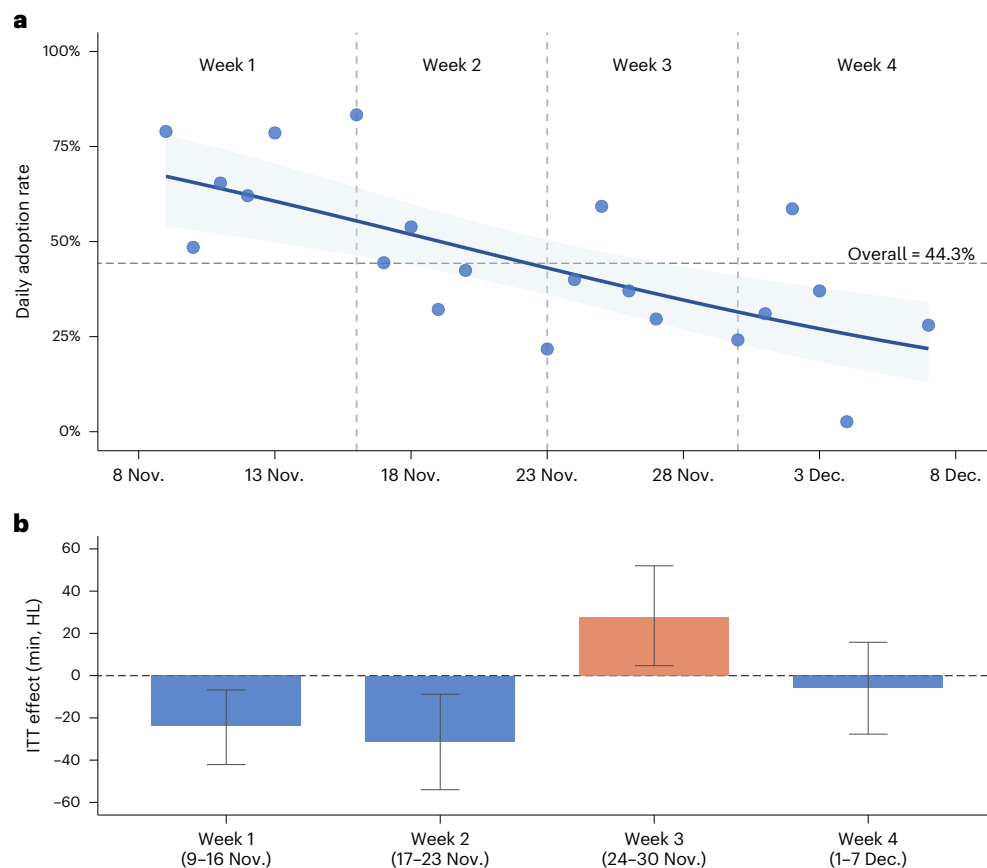


Fig. 3 | Adoption dynamics and temporal effect patterns. a, Daily SHAKED adoption rate in wing A. Points are observed daily proportions of wing A encounters in which physicians opened the SHAKED interface. The solid line is a logistic regression of daily adoption on calendar day, weighted by daily patient volume; the shaded band denotes the pointwise 95% confidence band of this fitted curve. The dashed horizontal line marks the overall adoption rate (44.3%; 259/584). Weekly adoption declined from 67.9% (week 1) to 29.7% (week 4). **b,** Weekly ITT effects on consultation cycle time. Bars show Hodges–Lehmann median shift estimates comparing all wing A with all wing B patients; negative

values indicate shorter consultation cycle time in wing A; error bars denote rank-based 95% CIs (Moses normal approximation to the Mann–Whitney statistic). Estimates with 95% CIs and two-sided *P* values from Mann–Whitney *U* tests: week 1, –24.0 min (95% CI –42.1 to –6.8; *P* = 0.006); week 2, –31.2 min (95% CI –54.0 to –8.8; *P* = 0.007); week 3 (coral), +27.5 min (95% CI 4.7 to 52.0; *P* = 0.018); and week 4, –5.9 min (95% CI –27.7 to 15.8; *P* = 0.607). *P* values are unadjusted for multiple comparisons; weekly estimates are exploratory (Supplementary Table 2). HL, Hodges–Lehmann.

Safety and expert validation

No adverse events were detected during the 4-week study period through any of the three surveillance mechanisms (real-time clinician reporting, post hoc log review or systematic chart review of 50 cases). Of 259 encounters in which physicians opened the SHAKED interface, 215 (83.0%) received complete system output, and 44 (17.0%) experienced delivery failures owing to server timeouts, data unavailability or network connectivity issues. No delivery failure resulted in an adverse clinical event. Post hoc analysis attributed failures primarily to Amazon Web Services (AWS) permission and network connectivity restrictions from the hospital network rather than infrastructure sensitivity to operational load. No substantial modifications were made to the SHAKED system during the study period. The system did not track recommendation acceptance or rejection; interaction logs recorded only binary interface access, precluding analysis of human–AI agreement patterns.

Expert validation of 100 consecutive SHAKED outputs showed differential diagnosis accuracy of 99% (99/100 outputs rated appropriate, 1 rated partial but not incorrect), test recommendation appropriateness of 100% (all recommendations within standard of care) and clinical summary plausibility of 100% (all summaries factually consistent with electronic health record (EHR) data). Interrater agreement before consensus was Gwet’s AC1 = 0.87 (95% CI 0.84 to 0.91); Cohen’s κ = 0.16, reflecting the prevalence paradox in high-agreement binary ratings (Supplementary Table 8).

Exploratory efficacy signal and post hoc exposure-level comparison

The primary outcome, ED length of stay, did not differ between wings (4.9 h in both wings, *P* = 0.99; Supplementary Table 4); a complete mapping of registered objectives and outcomes to the analyses, including the timing of and rationale for changes to the registered record, is provided in Supplementary Table 10.

ITT analysis comparing all wing A patients (*n* = 584) versus wing B controls (*n* = 554) was associated with shorter consultation time (–9.4 min, 95% CI –19.9 to 1.1, *P* = 0.077; Table 2 and Extended Data Fig. 2a). Per-protocol analysis comparing patients whose physicians opened SHAKED (*n* = 259) versus wing B controls showed –13.3 min (95% CI –26.2 to –0.4, *P* = 0.043). However, in wing A, users differed from non-users in radiology consultation frequency (84.2% versus 62.5%, SMD = 0.49), an imbalance attributable to selective adoption, instead of treatment effect, as confirmed by balanced radiology rates between wings (72.1% versus 73.6%, *P* = 0.601). After multivariable adjustment, the per-protocol estimate attenuated by 72%, whereas the ITT estimate was stable (11% attenuation), confirming selective adoption as a confounder in per-protocol analyses and supporting ITT as the less biased estimand (Supplementary Table 1 and Extended Data Fig. 1).

In a compliance-adjusted sensitivity analysis, the complier average causal effect (CACE) was –32.2 min (95% CI –65.2 to 0.9, *P* = 0.056). Anderson–Rubin confidence sets produced a virtually identical interval

Table 2 | Exploratory consultation cycle time estimates (ITT, per-protocol and CACE)

Analysis	Number	Effect (min)	95% CI	P
ITT	1,138 (584/554)	−9.4	−19.9 to 1.1	0.077
Per-protocol	259/554	−13.3	−26.2 to −0.4	0.043
CACE (Wald estimator)	1,138 (259 compliers)	−32.2	−65.2 to 0.9	0.056

CACE estimated via the Wald ratio estimator, presented as a post hoc instrumental variable sensitivity analysis relying on an untestable exclusion restriction. ITT and per-protocol estimates are Hodges–Lehmann median shifts with rank-based 95% CIs (Moses normal approximation to the Mann–Whitney statistic). Mean ITT difference (−14.3 min) exceeds Hodges–Lehmann shift owing to right-skewed distribution (skewness = 4.5). ITT and per-protocol P values are from two-sided Mann–Whitney U tests; all P values are unadjusted for multiple comparisons, and all estimates are exploratory.

(95% CI −65.6 to 0.9, $P = 0.057$). This estimate did not achieve statistical significance and should be interpreted as exploratory; the wide CI reflects the limited sample size and substantial between-patient heterogeneity (Supplementary Table 1). In exploratory ITT subgroup analyses, effects were directionally consistent across consultation types (radiology −10.3 min, $P = 0.081$; non-radiology −8.3 min, $P = 0.472$; interaction $P = 0.35$) and larger among admitted patients (−15.8 min, $P = 0.047$) than discharged patients (−4.0 min, $P = 0.564$; interaction $P = 0.52$). Treatment effects were also similar between female and male patients (female −11.5 min, $P = 0.14$; male −8.3 min, $P = 0.25$; sex-by-allocation interaction $P = 0.88$).

In a post hoc analysis, to distinguish treatment effect from selection artifact, we stratified outcomes by exposure level in wing A, using wing B controls as common reference (Table 3 and Extended Data Fig. 3). Patients with passive exposure (system autogenerated output without physician interface access) showed consultation times statistically indistinguishable from controls: median 179.1 min versus 175.4 min in wing B (Hodges–Lehmann shift −1.3 min, 95% CI −14.2 to 12.3, $P = 0.84$). Patients whose physicians initiated SHAKED requests but received no output owing to system failure showed no consultation time difference: median 154.0 min (shift −11.5 min, 95% CI −39.1 to 20.9, $P = 0.45$; $n = 44$). Patients whose physicians opened SHAKED and received full output showed shorter consultation times: median 158.3 min (shift −13.6 min, 95% CI −26.4 to −0.4, $P = 0.048$). The three-level pattern (passive null, technical failure null and active engagement significant) is suggestive of, but does not establish, a content-dependent association requiring deliberate physician interaction with AI-generated content.

During week 3, wing A consultation times increased (median 195.8 min), whereas wing B remained stable (median 166.7 min). Wing B showed no temporal trend across all 21 study days ($P = 0.920$). Excluding week 3 yielded −20.1 min ($P < 0.001$) for ITT and −24.2 min ($P = 0.001$) for per protocol, compared with −9.4 min ($P = 0.077$) and −13.3 min ($P = 0.043$) in the full sample. We retained week 3 in the primary analysis to avoid post hoc selection bias (Supplementary Table 2).

Discussion

In this prospective controlled pilot study, we evaluated a multimodel LLM-based CDSS during live deployment in a tertiary ED, following the four DECIDE-AI stage 1 objectives: preliminary evidence of clinical utility, safety profile characterization, human factors assessment and preparation for larger summative trials. The study was designed to generate point estimates and variance parameters for powering a definitive trial, not to provide confirmatory efficacy evidence. Three concurrent surveillance mechanisms identified no adverse events across 584 intervention-arm encounters, and expert review was clinically accurate. The implementation dynamics findings deserve particular attention: adoption decayed from 67.9% to 29.7% over 4 weeks, with workload-sensitive disengagement and preferential engagement

Table 3 | Post hoc exposure-level comparison: mechanism of effect

Exposure Level	Number	Effect (min)	95% CI	P
Wing B controls (ref)	554	–	–	–
Passive exposure	273	−1.3	−14.2 to 12.3	0.84
Technical failure	44	−11.5	−39.1 to 20.9	0.45
Active engagement	215	−13.6	−26.4 to −0.4	0.048

Hodges–Lehmann shift estimates versus wing B controls. An additional 52 wing A patients with no SHAKED-processed data (no exposure) showed shorter consultation times than controls (−29.7 min, $P = 0.018$) and were excluded from the exposure-level comparison because no AI content was generated for these encounters. Estimates are Hodges–Lehmann median shifts with rank-based 95% CIs (Moses normal approximation); P values are from two-sided Mann–Whitney U tests versus wing B controls, unadjusted for multiple comparisons. Exposure groups are postrandomization classifications; estimates are exploratory.

for radiology consultations representing barriers rarely documented prospectively in studies of AI-based clinical decision support.

Adoption declined steadily across the study period, matching temporal decay patterns that implementation science describes but rarely documents prospectively in AI-based clinical tools^{11,12,21}. Engagement declined with later ED arrival hour, meaning clinicians disengaged during the highest workload periods^{21–23}. Physicians were markedly more likely to use SHAKED for radiology consultations than for other types, despite no differential effectiveness, suggesting engagement driven by perceived utility rather than measured effectiveness across consultation types^{24,25}. Several factors may have contributed to the observed decay. The 17% delivery failure rate may have eroded clinician trust, consistent with evidence that CDSS reliability thresholds influence sustained adoption²². Additionally, the absence of feedback on recommendation accuracy may have limited perceived value, and the opt-in activation model required physicians to interrupt their workflow during periods of increasing cognitive load.

These patterns point to specific design changes for future deployments: default-active opt-out architectures, reduced friction interfaces for high-workload periods and structured re-engagement protocols to counter novelty fatigue²⁶. More broadly, these findings illustrate that deployment success is determined not by algorithmic capability alone but by the surrounding implementation ecosystem: infrastructure reliability, workflow integration and organizational context^{27,28}.

Embedded effectively, an LLM-based decision-support layer could reshape the ED consultation pathway: by pregenerating a synthesized history and candidate workup before the physician opens the record, it could shift clinician effort from information assembly toward decision-making and enable earlier, better-specified consultation requests (a hypothesis for future testing). Realizing that change, however, depends less on incremental model accuracy than on embedding the tool at the points of lowest workflow friction and sustaining engagement through the high-workload periods when it is currently abandoned.

We examined the exposure-level comparison to distinguish active engagement from non-specific effects, addressing the question posed of whether LLM-based systems require deliberate physician interaction or can improve outcomes through passive availability alone^{18,19}. Passive exposure, where the system generated output but the physician did not engage, produced no time reduction, arguing against both Hawthorne effects²⁹ and training-alone explanations, as these patients were treated by the same trained physicians but without active SHAKED engagement.

Technical failures, where physicians intended to use the system but received no AI content, also showed no benefit, serving as an internal control for physician self-selection, although the small

sample ($n = 44$) limits power. Only active engagement was associated with shorter consultation time, suggesting a mechanism requiring deliberate physician interaction rather than allocation awareness alone. Four observations argue against pure selection bias: (1) the null technical failure result, (2) the null passive exposure result, (3) equivalent performance of wing A non-users versus wing B controls on all measured outcomes (Supplementary Table 6) and (4) stability of the ITT estimate after covariate adjustment, in contrast to the marked attenuation of the per-protocol estimate (Supplementary Table 1). Although no single test is definitive, their convergence is consistent with a mechanism requiring active physician engagement with AI-generated content.

The registered primary outcome, ED length of stay, did not differ between wings, so the consultation cycle time results that follow are best read as a prespecified but exploratory efficacy signal. ITT analysis showed a directional, non-significant reduction in consultation cycle time, and a post hoc instrumental variable sensitivity analysis was directionally consistent; neither reached statistical significance. Both should therefore be read as effect-size estimates to inform sample size calculations for a definitive trial, not as evidence of efficacy. No established minimal clinically important difference exists for ED consultation cycle time³; published interventions achieved reductions of 21 to 92 min^{30,31}, providing context for interpreting our estimates. These findings expand on refs. 1718, in which LLM capability in ED clinical decision support was evaluated, by providing prospective evidence from live clinical deployment with concurrent measurement of adoption and efficacy.

Expert review supported the clinical appropriateness of system outputs (Supplementary Table 8). No adverse events were detected, although the study was not powered for rare harms or diagnostic anchoring. The observed delivery success rate, while acceptable for a pilot, falls below the reliability threshold likely required for sustained adoption at scale. Potential biases in LLM outputs toward specific populations require evaluation in larger, more diverse cohorts.

Strengths of this evaluation include its prospective design during live clinical deployment, the concurrent measurement of adoption and efficacy, and an exposure-level comparison that distinguishes active engagement from passive availability.

This study has several limitations. This was a single-center, 4-week pilot study; generalizability is limited accordingly. Although allocation balance was confirmed through four convergent tests (Supplementary Table 6), unmeasured confounding cannot be excluded³². Clinician training included consultation efficiency content independent of SHAKED, introducing potential expectation effects; however, the null passive exposure result argues against a training-alone explanation, as patients treated by trained physicians who did not engage with SHAKED showed no consultation time improvement. The absence of physician identifiers precluded individual adoption analysis. Consultation cycle time is a process measure^{33,34}; ED length of stay, a primary outcome specified in the approved study protocol, did not differ between wings (Supplementary Table 4), although the study was underpowered to detect downstream throughput differences. Indirect contamination could have attenuated ITT estimates. The CACE likely represents an upper bound during the novelty adoption phase.

These findings inform randomized trial design but do not justify clinical deployment. On the basis of the observed effect size and variance, a definitive trial would require approximately 225 compliers per arm to achieve 80% power (Supplementary Table 5). Required design elements include physician- or shift-level randomization to minimize contamination³⁵, minimum 6-month duration to capture learning curves and seasonal variation, and stepped-wedge design for operational advantages³⁶. The outcome structure should include consultation time as the primary endpoint with secondary endpoints including ED length of stay, diagnostic error rates, disposition

appropriateness and clinician cognitive load. Future trial instrumentation should additionally capture recommendation-level uptake (clinician review, acceptance or override), which the present interface logs did not record, and incorporate independent validation of the consultation timestamps as a prespecified data-quality check; a target trial emulation framework would further address the time-varying confounding introduced by dynamic adoption. Implementation optimization should address the adoption barriers identified in this study: novelty fatigue through periodic re-engagement training, activation friction through default-active architecture and delivery reliability through infrastructure hardening, with fidelity monitored using established frameworks²⁶.

This DECIDE-AI stage 1 evaluation showed that a multimodel LLM-based CDSS could be integrated into a live tertiary ED with no adverse events observed, 99% differential diagnosis accuracy and 100% test recommendation appropriateness on expert review. Exposure-level comparison showed no benefit from passive system availability, whereas active engagement was associated with shorter consultation time, pointing to a mechanism requiring deliberate physician interaction and not background presence alone.

However, adoption decayed from 67.9% to 29.7% over 4 weeks, suggesting that sustained clinician engagement, rather than diagnostic capability, is likely the principal barrier for clinical AI in emergency medicine. These compliance rates and effect estimates now provide the foundation for a definitive randomized trial with physician-level randomization, extended duration and patient-centered outcomes.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-026-04601-5>.

References

- Hoot, N. R. & Aronsky, D. Systematic review of emergency department crowding: causes, effects, and solutions. *Ann. Emerg. Med.* **52**, 126–136 (2008).
- Morley, C., Unwin, M., Peterson, G. M., Stankovich, J. & Kinsman, L. Emergency department crowding: a systematic review of causes, consequences and solutions. *PLoS ONE* **13**, e0203316 (2018).
- Voaklander, B. et al. Interventions to improve consultations in the emergency department: a systematic review. *Acad. Emerg. Med.* **29**, 1475–1495 (2022).
- Brick, C. et al. The impact of consultation on length of stay in tertiary care emergency departments. *Emerg. Med. J.* **31**, 134–138 (2014).
- Piliuk, K. & Tomforde, S. Artificial intelligence in emergency medicine: a systematic literature review. *Int. J. Med. Inform.* **180**, 105274 (2023).
- Grant, K., McParland, A., Mehta, S. & Ackery, A. D. Artificial intelligence in emergency medicine: surmountable barriers with revolutionary potential. *Ann. Emerg. Med.* **75**, 721–726 (2020).
- Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
- Plana, D. et al. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw. Open* **5**, e2233946 (2022).
- Han, T. et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit. Health* **6**, e367–e373 (2024).
- Oikonomidi, T. et al. Applications of artificial intelligence for real-world evidence generation: a protocol for a living scoping review. *BMJ Open* **16**, e109725 (2026).

11. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
12. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
13. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
14. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
15. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
16. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
17. Williams, C. Y. K., Miao, B. Y., Kornblith, A. E. & Butte, A. J. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nat. Commun.* **15**, 8236 (2024).
18. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
19. Preiksaitis, C. & Rose, C. The role of large language models in transforming emergency medicine: scoping review. *JMIR Med. Inform.* **12**, e53787 (2024).
20. Gilboy, N., Tanabe, T., Travers, D. & Rosenau, A. M. *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4*. AHRQ Publication No. 12-0014 (AHRQ, 2011).
21. Greenhalgh, T. et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J. Med. Internet Res.* **19**, e367 (2017).
22. R Core Team R: *A Language and Environment for Statistical Computing* (R Core Team, 2024).
23. 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).
24. Kawamoto, K., Houlihan, C. A., Balas, E. A. & Lobach, D. F. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* **330**, 765 (2005).
25. Roshanov, P. S. et al. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. *BMJ* **346**, f657 (2013).
26. Proctor, E. K. et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm. Policy Ment. Health* **38**, 65–76 (2011).
27. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* **91**, 444–455 (1996).
28. Kwong, J. C. C., Nickel, G. C., Wang, S. C. Y. & Kvedar, J. C. Integrating artificial intelligence into healthcare systems: more than just the algorithm. *npj Digit. Med.* **7**, 52 (2024).
29. McCambridge, J., Witton, J. & Elbourne, D. R. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J. Clin. Epidemiol.* **67**, 267–277 (2014).
30. Geskey, J. M., Geeting, G., West, C. & Hollenbeak, C. S. Improved physician consult response times in an academic emergency department after implementation of an institutional guideline. *J. Emerg. Med.* **44**, 999–1006 (2013).
31. Soong, C., High, S., Morgan, M. W. & Ovens, H. A novel approach to improving emergency department consultant response times. *BMJ Qual. Saf.* **22**, 299–305 (2013).
32. VanderWeele, T. J. & Ding, P. Sensitivity analysis in observational research: introducing the E-value. *Ann. Intern. Med.* **167**, 268–274 (2017).
33. Donabedian, A. The quality of care: how can it be assessed?. *JAMA* **260**, 1743–1748 (1988).
34. Mant, J. Process versus outcome indicators in the assessment of quality of health care. *Int. J. Qual. Health Care* **13**, 475–480 (2001).
35. Torgerson, D. J. Contamination in trials: is cluster randomisation the answer? *BMJ* **322**, 355–357 (2001).
36. Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J. & Lilford, R. J. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* **350**, h391 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2026

Methods

Study design

We conducted a prospective controlled pilot study evaluating SHAKED in a live ED setting from 9 November to 7 December 2025 (4 weeks, 21 study dates), designed and reported according to DECIDE-AI guidelines¹³ (Supplementary Information) and STROBE checklist for observational studies (Supplementary Document 3). Trial registration: ClinicalTrials.gov [NCT06902675](https://clinicaltrials.gov/ct2/show/study/NCT06902675); see Supplementary Table 11 for the registration version history. Patients and the public were not involved in the design, conduct or reporting of this study.

Setting

The study was performed at Rambam Health Care Campus, a tertiary-care ED comprising two parallel clinical units (wing A, intervention; wing B, control) with comparable staffing, volumes and case mix.

Eligibility criteria

All adult patients (≥ 18 years) presenting during weekday morning shifts (07:00 to 15:00) were included. Patients without specialty consultations were excluded from time analyses ($n = 8$, $<1\%$).

Intervention and controls

The SHAKED system was deployed as part of routine ED operations before the study period; the system was already in routine use and was not modified for the study.

SHAKED is a CDSS using a multimodel architecture orchestrating several specialized LLMs^{13,14} (Claude Sonnet 4.5, Claude Haiku 4.5, Amazon Nova Lite and Amazon Nova Pro) with retrieval-augmented generation³⁷ drawing from $>40,000$ clinical documents. Each specialized model performed a defined role in a sequential pipeline as opposed to an ensemble or voting scheme; all models were accessed as fixed, version-pinned endpoints through AWS Bedrock, and no model was fine-tuned or modified during the pilot. The system functioned exclusively as physician-supervised decision support^{38,39}, delivering medical history synthesis, clinical decision support (differential diagnoses, workup recommendations, consultation content) and a conversational interface for querying patient records and clinical literature. Each clinical statement in these outputs was linked to its source, whether an item in the patient's record or a retrieved literature passage, so that physicians could trace every assertion to its origin. Input data were extracted automatically from the EHR; no patient data were used to train the models. The system operated through a two-trigger protocol: assignment of a patient to wing A initiated automated extraction from the EHR, and completion of triage documentation triggered automatic generation of the medical history synthesis and decision support content, so that output was available before the physician encounter began; a conversational module additionally allowed ad hoc queries. Physicians accessed SHAKED directly from the Prometheus record through a dedicated per-patient control that opened the patient's already-generated case in a companion web application. All outputs required explicit physician review; the system had no autonomous decision-making or order initiation capability (Fig. 2 and Extended Data Fig. 4).

Wing B operated as the standard-of-care control unit without SHAKED access; system logs confirmed that no physician accessed the SHAKED interface during a wing B encounter.

Physicians who treated wing A patients completed predeployment training that included consultation efficiency content delivered independently of SHAKED; the same physicians also treated wing B patients during routine rotation.

Allocation procedures

At the triage registration station, a preprinted roll of sequential stickers (strictly alternating A, B, A, B) assigned each patient to a wing at the moment of registration, before any clinical assessment, electronic

record entry or physician contact. The charge nurse was blinded to SHAKED intervention status.

Outcome measures

Outcomes are designated registered primary, prespecified or post hoc. Under the DECIDE-AI stage 1 framework, the prespecified primary objectives were adoption and implementation. The registered primary outcome (ClinicalTrials.gov [NCT06902675](https://clinicaltrials.gov/ct2/show/study/NCT06902675)) is ED length of stay, reported plainly, including its null result (4.9 h in both wings, $P = 0.99$). Consultation cycle time was the prespecified principal exploratory efficacy endpoint, with the ITT comparison as the primary estimand; it is defined as the interval between two EHR-generated timestamps: electronic consultation order entry (start) and specialist recommendation documentation (stop)^{40,41}, analyzed through three estimands: ITT, per protocol and CACE using ward allocation as an instrumental variable^{27,42–46}. Wing A patients were classified into four exposure levels: no exposure, passive exposure, technical failure and active engagement, defined as physician-initiated interface access (Extended Data Fig. 3). Radiology consultation was defined as any imaging study requiring radiologist interpretation; bedside imaging not requiring formal interpretation was excluded. The exposure-level classification and the per-protocol and CACE estimates are reported as post hoc analyses; adoption rate and usage patterns are prespecified implementation outcomes.

Data sources and how data were collected

All clinical data were extracted from the Prometheus EHR for the two ED units as two source extracts, one per wing, each containing the patient record, diagnoses, consultation orders and responses, and, for wing A, the SHAKED usage logs. Patient demographics, triage acuity, disposition, diagnoses, and the consultation order and specialist recommendation timestamps were taken directly from automated EHR fields rather than manual entry. For each wing A case, the system parsed the raw record into a deduplicated, chronological event list in which every event received a unique reference identifier, providing the source citations carried by each output. SHAKED usage was recorded by the system logs and linked to each patient by case number, yielding two flags, interface access and content delivery, from which encounters were classified as passive exposure, technical failure or active engagement. A single documented cleaning pipeline filtered each extract to the pilot window, derived the usage flags, deduplicated cross-wing patient moves and computed the outcome variables; structured field completeness exceeded 99%.

Expert review procedures

Safety was monitored through real-time clinician reporting, post hoc log review and systematic chart review; 100 consecutive outputs were independently reviewed by four blinded expert physicians. For each case, reviewers had access to the SHAKED output together with the patient's triage-recorded reason for presentation and prior clinical history. The four reviewers were two neurologists and two emergency medicine specialists, blinded to one another's ratings; each output was scored on five binary (valid or not valid) items spanning differential diagnosis, recommended workup and medical history summary accuracy. Assessment relied on clinical judgment; no standardized numerical rubric was used, and disagreements were not formally adjudicated, so raw agreement rates are reported. Expert-rated differential diagnosis accuracy was 99% (99 of 100), with 100% recommended test appropriateness and 100% clinical summary plausibility. Interrater agreement across the 500 paired ratings was 88.8% observed agreement, with a Gwet AC1 of 0.87 (95% CI 0.84 to 0.91); the Cohen's κ on the same ratings was 0.16, deflated by the high prevalence of valid ratings. We report this as interrater agreement, not as validation.

Statistical analysis

All statistical analyses were performed in Python; R was used only for figure rendering. ITT was the primary estimand for consultation

cycle time; because the distribution was right-skewed, location shifts were estimated as Hodges–Lehmann median shifts with rank-based 95% CIs (Moses normal approximation), and cluster robust standard errors (clustered by arrival date, 21 clusters) accounted for within-day correlation. The registered primary outcome, ED length of stay, was compared between wings with its minimum detectable effect reported. Per-protocol and CACE estimates were computed as post hoc sensitivity analyses, the latter using wing allocation as an instrument under an untestable exclusion restriction. Multiplicity across a family of 32 tests was controlled by the Holm–Bonferroni procedure, and *E* values characterized sensitivity to unmeasured confounding. No formal sample size calculation was performed; supporting a DECIDE-AI stage 1 pilot, all consecutive eligible patients presenting during the fixed 4-week window were analyzed.

Statistical methods, exposure classification details, clinician training procedures and allocation balance verification are described in the Supplementary Information^{22,47}.

Ethics

This study was approved by the Institutional Review Board (Helsinki Committee) of Rambam Health Care Campus (approval number 0026-24RMB) and was conducted in accordance with the principles of the Declaration of Helsinki.

The SHAKED system was operating as part of routine ED care before the study began; no intervention was introduced or modified for research purposes. The system's working parameters, including the settings defined across wing A and wing B in the ED and the prespecified operational objectives of adoption and execution (Supplementary Table 10), were established prospectively by our institutional governance committee, comprising hospital leadership. These decisions constituted institutional implementation and operational choice, not a research intervention, and the study itself is a retrospective analysis of the clinical data generated during that routine implementation.

All information analyzed had been generated during routine clinical care; data were extracted from the hospital EHR only after the relevant encounters had concluded and were analyzed in de-identified form. All direct and indirect patient identifiers were removed, and no individual physician identifiers were recorded, so no identifiable data were examined. On this basis, the Israeli Ministry of Health granted a waiver of informed consent under Director General's Circular number 1/2018 (Supplementary Document 2), which defines secondary use as the use of health information outside the scope of direct clinical care (section 4.4).

The study was registered on ClinicalTrials.gov ([NCT06902675](https://clinicaltrials.gov/ct2/show/study/NCT06902675)).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data underlying this study are released as a de-identified, patient-level analytic file accompanied by a data dictionary that defines every field, its units and its coding. The file contains the variables required to reproduce the reported analyses. All direct identifiers, exact timestamps and free-text fields (including diagnosis text) have been removed, and clinician identifiers are not included. The publicly deposited analysis code reproduces the reported results directly from this file, so the file keeps the integer age and the real (unshifted) visit date that the code needs. For that reason, and in accordance with Israeli health privacy law and the study's ethics approval, it is shared under controlled access rather than as an open-release file. Requests should be made in writing to the corresponding author, S.S. (shahar.shell@technion.ac.il), and are reviewed and approved by the corresponding author together with the Rambam Health Care Campus Institutional

Review Board. Access is granted to applicants who (1) are affiliated with an accredited academic, medical or non-profit research institution able to enter into a data-use agreement; (2) submit a written analysis plan stating a specific scientific question, to verify or extend the published analyses; (3) hold or obtain ethics approval or exemption from their own institution; and (4) sign a data-use agreement prohibiting re-identification, onward transfer and commercial use, and restricting use to the stated purpose. Receipt of each request will be acknowledged within 14 days, and a decision is typically communicated within 14 days of receipt of a complete application. Controlled access will be provided for a minimum of 5 years following publication; after this period, the dataset remains under Rambam institutional governance, with requests considered under the same procedure.

Code availability

The analysis and figure-generation code reproducing the reported quantitative results is publicly available under the MIT license at <https://github.com/llironlibo/SHAKED-analysis> and permanently archived at Zenodo (<https://doi.org/10.5281/zenodo.20736931>)⁴⁸. The repository holds the full analysis and figure-generation pipeline, with a top-level reproduction entry point, a README and a requirements file, and it contains no patient data. The data preparation script that reads the raw protected records is included for transparency, but it cannot be run without the controlled access raw files.

References

- Lewis P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. 34th International Conference on Neural Information Processing Systems (NIPS '20)* (eds Larochelle, H. et al.) 9459–9474 (Curran, 2020).
- Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
- Ethics and Governance of Artificial Intelligence for Health: WHO Guidance* (WHO, 2021).
- Shin, S., Lee, S. H., Kim, D. H. & Kim, K. H. The impact of the improvement in internal medicine consultation process on emergency department length of stay. *Am. J. Emerg. Med.* **36**, 620–624 (2018).
- Ravi, A., Shochat, G., Wang, R. C. & Khanna, R. Improvements to emergency department length of stay and user satisfaction after implementation of an integrated consult order. *J. Am. Coll. Emerg. Physicians Open* **4**, e12922 (2023).
- Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28**, 3083–3107 (2009).
- Greenland, S. An introduction to instrumental variables for epidemiologists. *Int. J. Epidemiol.* **29**, 722–729 (2000).
- Staiger, D. & Stock, J. H. Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–586 (1997).
- Hodges, J. L. Jr & Lehmann, E. L. Estimates of location based on rank tests. *Ann. Math. Stat.* **34**, 598–611 (1963).
- Meunier, P. Y., Raynaud, C., Guimaraes, E., Gueyffier, F. & Letrillart, L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. *Ann. Fam. Med.* **21**, 57–69 (2023).
- Python v.3.13 (Python Software Foundation, 2024).
- llironlibo. llironlibo/SHAKED-analysis: v1.1.0. Zenodo <https://doi.org/10.5281/zenodo.20736931> (2026).

Acknowledgements

We thank the physicians, nursing staff and administrative personnel of the Rambam Health Care Campus Emergency Department for their participation and for sustaining the pilot through routine clinical operation. Rambam Health Care Campus (Haifa, Israel)

Emergency Department served as the clinical deployment site for SHAKED, providing the live ED setting, EHR integration, clinical leadership and physician participation. We further acknowledge institutional support from Rambam including clinical infrastructure, EHR integration, a protected computing environment and research personnel.

Author contributions

S.S. conceived and supervised the study and is the corresponding author. L.L. led the data analysis, developed the analysis code and prepared the figures. S.S. and L.L. wrote the original draft. D.A. contributed to the computational methodology and the modeling approach. A.A., A.G., M.S. and K.M. contributed to clinical implementation and data collection. All authors reviewed the analyses, contributed to revising the manuscript and approved the final version.

Funding

This work was supported by AWS Partner funding and \$20,000 in cloud computing credits, which enabled access to secure hospital-grade infrastructure, the AWS Bedrock AI platform and scalable LLM inference for SHAKED's clinical deployment. The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Beyond the AWS support

disclosed above, this research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Additional information

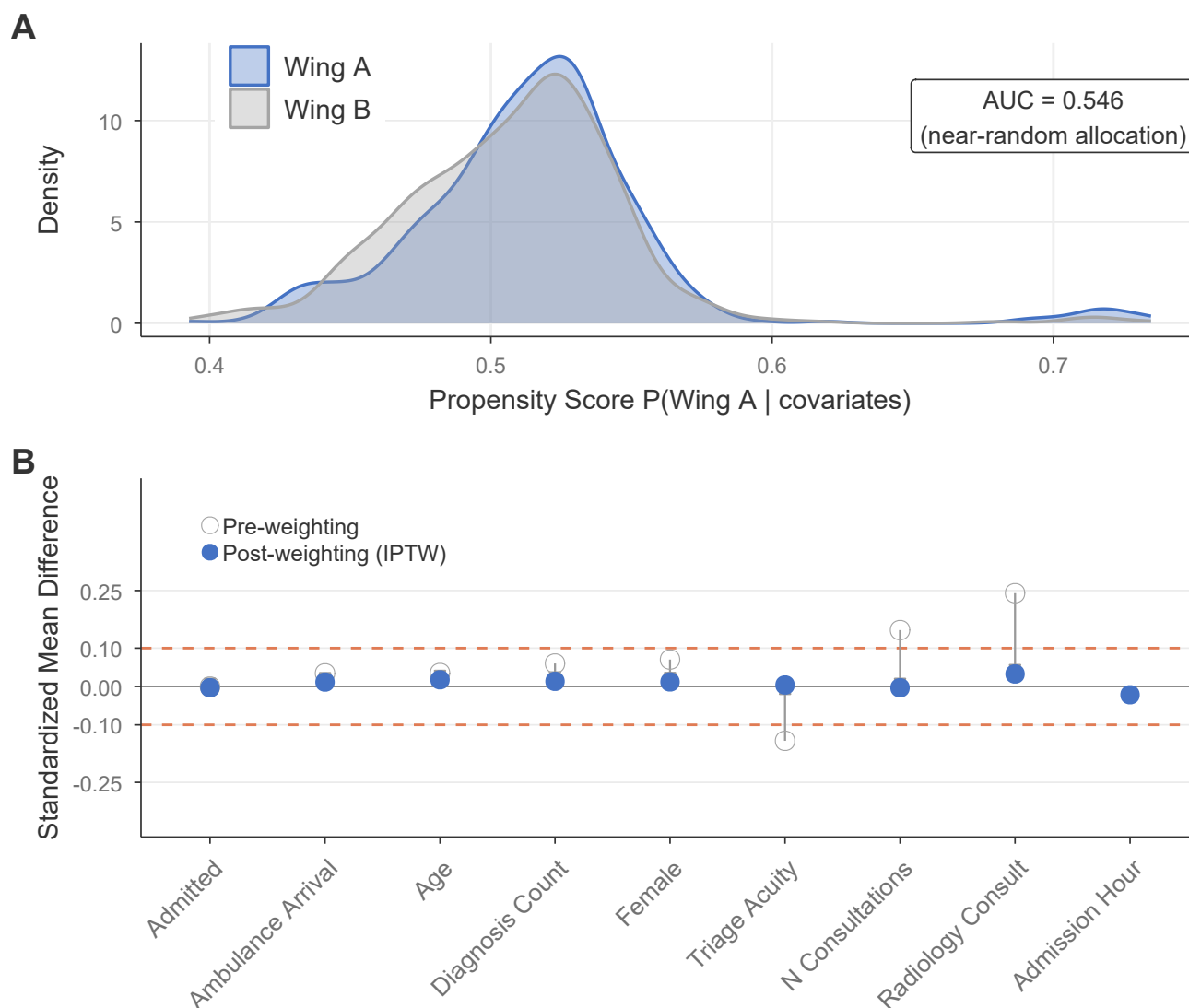
Extended data is available for this paper at <https://doi.org/10.1038/s41591-026-04601-5>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-026-04601-5>.

Correspondence and requests for materials should be addressed to Shahar Shelly.

Peer review information *Nature Medicine* thanks Myura Nagendran, Zhongheng Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Mattia Andreoletti and Ming Yang, in collaboration with the *Nature Medicine* team. Peer reviewer reports are available.

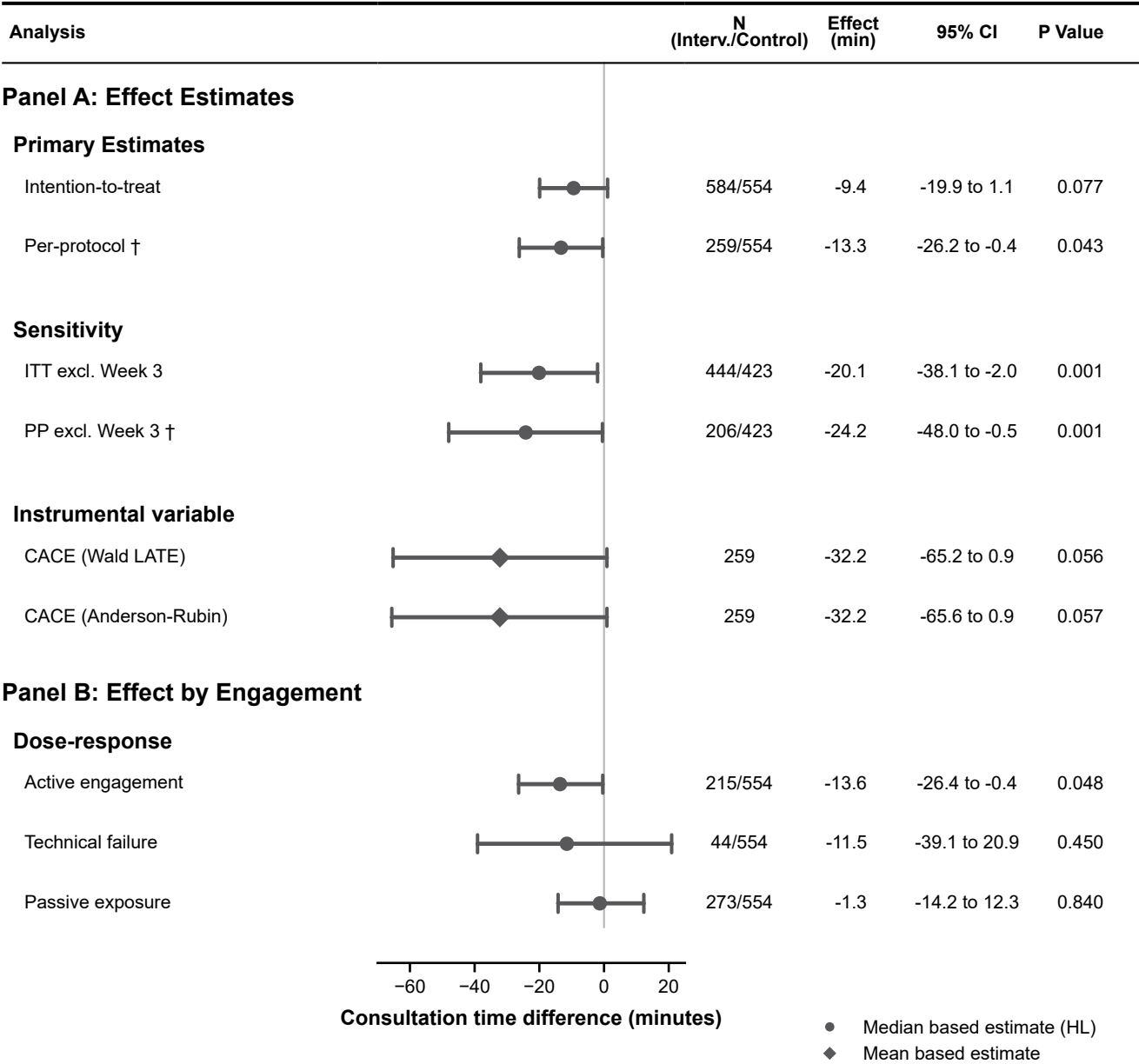
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Propensity Score Distribution and Covariate Balance.

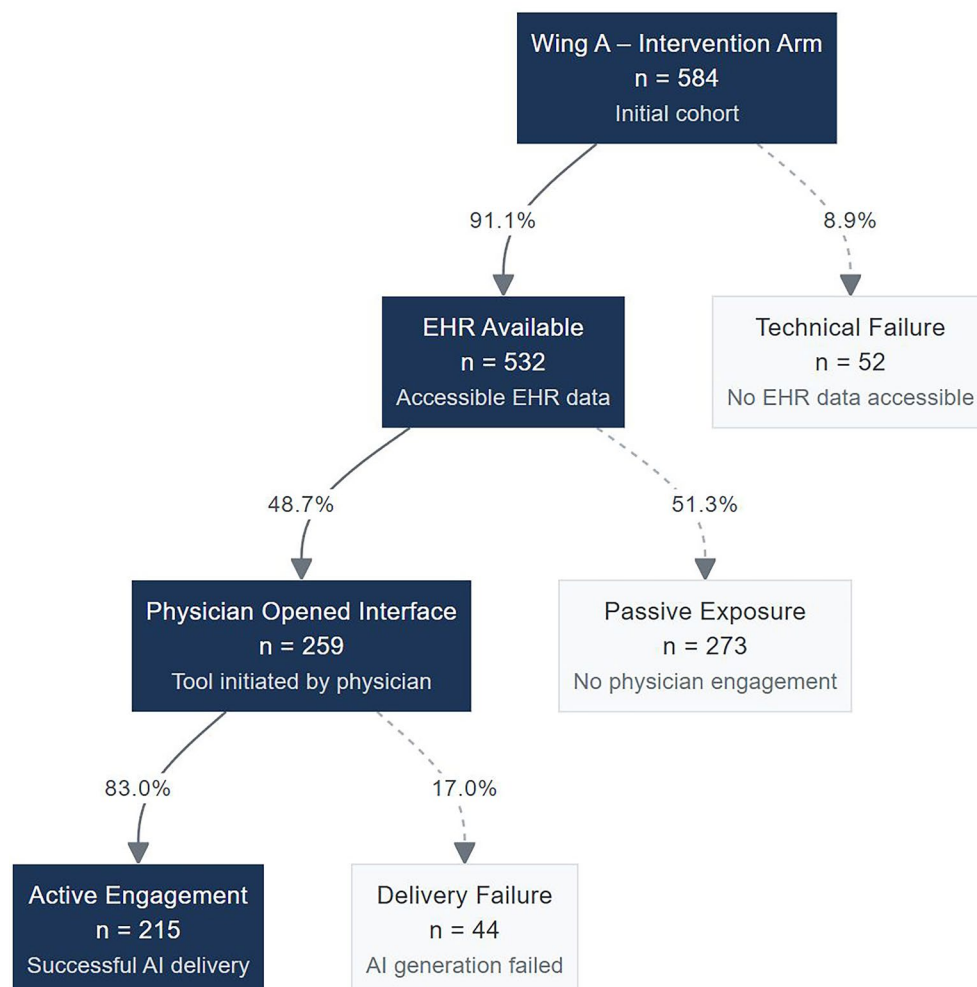
Abbreviations: ATE, average treatment effect; AUC, area under the curve; IPTW, inverse probability of treatment weighting; SMD, standardized mean difference. **(A)** Distribution of propensity scores (predicted probability of wing assignment) for Wing A (blue) and Wing B (gray), estimated by logistic regression including age, sex, triage acuity (P1–P5), radiology consultation, admission status, and

diagnosis count ($n = 1,133$; AUC, 0.543). **(B)** SMD love plot showing covariate balance before (open circles) and after (filled circles) IPTW for the per-protocol comparison (SHAKED users vs Wing B controls). Dashed vertical lines indicate ± 0.10 threshold. All post-weighting SMDs fall within ± 0.10 (maximum, 0.032). Propensity score model: lasso-logistic regression with stabilized ATE weights truncated at 1st and 99th percentiles.



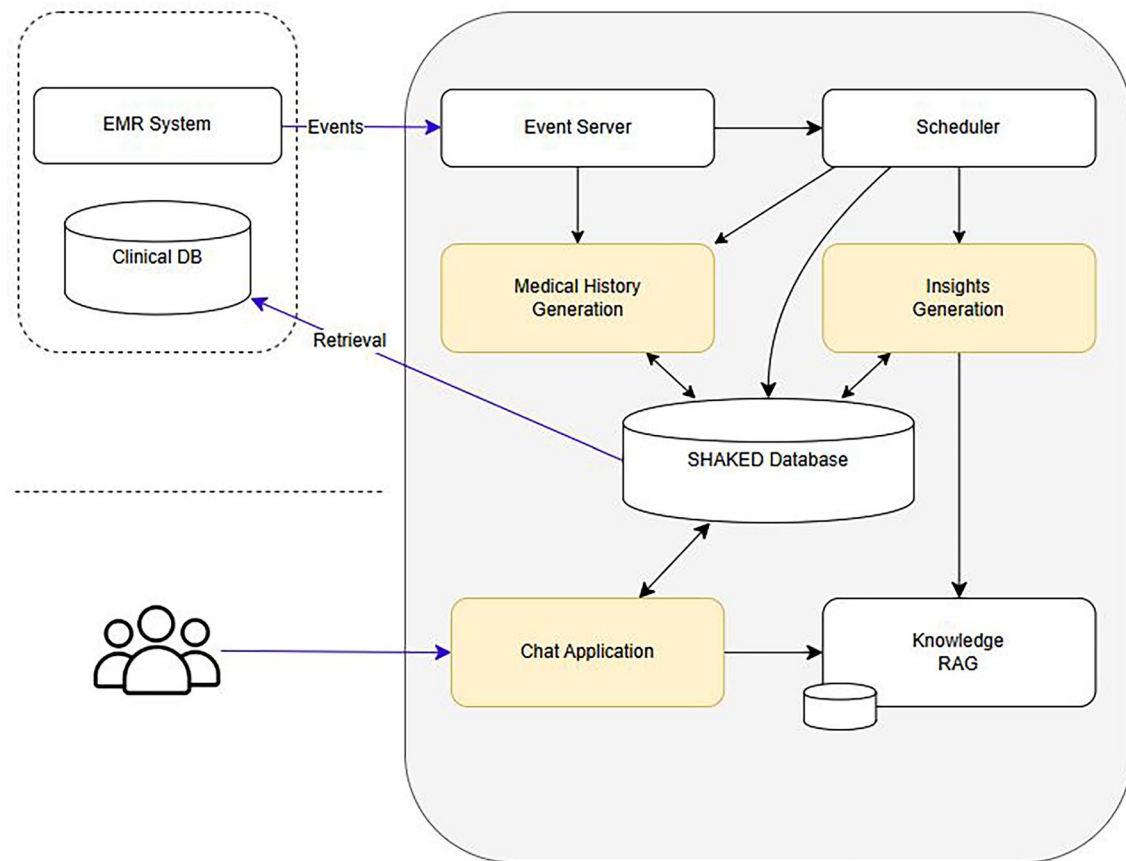
Extended Data Fig. 2 | Association of SHAKED decision support with consultation cycle time. (A) Forest plot of consultation cycle time differences (minutes). Circles denote Hodges-Lehmann median shift estimates with rank-based 95% confidence intervals (CIs; Moses normal approximation to the Mann-Whitney statistic); p-values are from two-sided Mann-Whitney U tests. Diamonds denote mean-based complier average causal effect (CACE) estimates

from instrumental variable analysis (Wald estimator; two-sided). (B) Dose-response analysis comparing three Wing A engagement levels against Wing B controls; Hodges-Lehmann estimates with rank-based 95% CIs, two-sided Mann-Whitney U tests. All p-values are unadjusted for multiple comparisons and all estimates are exploratory. Per-protocol (PP) estimates may be subject to selective adoption bias (Supplementary Table 1).



Extended Data Fig. 3 | Mechanism-Based Exposure Groups Within the Intervention Arm. Abbreviations: AI, artificial intelligence; EHR, electronic health record. Patients allocated to Wing A (intervention, n = 584) were classified into four mutually exclusive exposure groups based on three sequential conditions: EHR data

availability for system processing, physician activation of the SHAKED interface, and successful AI output delivery. Percentages on arrows indicate the proportion of patients from the preceding node transitioning to each group.



Extended Data Fig. 4 | SHAKED System Architecture. Clinical events from the EMR trigger an event-driven pipeline coordinated by an Event Server and Scheduler. Patient data are retrieved from the Clinical Database to generate structured medical histories and automated clinical insights, which are stored

in the central SHAKED Database. Clinicians access the system through a chat interface, supported by a Retrieval-Augmented Generation (RAG) knowledge module for evidence-based responses. Arrows indicate data flow within the secure SHAKED environment.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Electronic health record system (Rambam Health Care Campus) all clinical data were extracted from automated EHR fields. SHAKED clinical decision support system, built on AWS Bedrock using Claude Sonnet 3.5, Claude Haiku 3.5, Amazon Nova-lite, and Amazon Nova Pro.
Data analysis	https://doi.org/10.5281/zenodo.20736931 Repository holds the full analysis and figure-generation pipeline

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The data underlying this study are released as a de-identified, patient-level analytic file accompanied by a data dictionary that defines every field, its units, and its coding. The file contains the variables required to reproduce the reported analyses. All direct identifiers, exact timestamps, and free-text fields (including diagnosis text) have been removed, and clinician identifiers are not included. The publicly deposited analysis code reproduces the reported

results directly from this file, so the file keeps the integer age and the real (unshifted) visit date that the code needs. For that reason, and in accordance with Israeli health-privacy law and the study's ethics approval, it is shared under controlled access rather than as an open-release file. Requests should be made in writing to the corresponding author, Prof. Shahar Shelly (s_shelly@rambam.health.gov.il), and are reviewed and approved by the corresponding author together with the Rambam Health Care Campus Institutional Review Board. Access is granted to applicants who (i) are affiliated with an accredited academic, medical, or non-profit research institution able to enter into a data-use agreement; (ii) submit a written analysis plan stating a specific scientific question, to verify or extend the published analyses; (iii) hold or obtain ethics approval or exemption from their own institution; and (iv) sign a data-use agreement prohibiting re-identification, onward transfer, and commercial use, and restricting use to the stated purpose. Receipt of each request will be acknowledged within 14 days, and a decision is typically communicated within 14 days of receipt of a complete application. Controlled access will be provided for a minimum of five years following publication; after this period the dataset remains under Rambam institutional governance, with requests considered under the same procedure.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Use the terms *sex* (biological attribute) and *gender* (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.

Reporting on race, ethnicity, or other socially relevant groupings

Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

Identify the organization(s) that approved the study protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	1,138 - No a priori calculation; all consecutive eligible patients in the four-week window (DECIDE-AI Stage 1). Observed estimates power the definitive trial (Supplementary Table 5).
Data exclusions	Eight of 1,146 patients (<1%) were excluded for having no specialty consultation (pre-established criterion); no other exclusions.
Replication	Single-site DECIDE-AI Stage 1 pilot embedded in routine care; findings inform a definitive multi-site trial.
Randomization	Strict alternating assignment (A/B/A/B) at triage; balance verified (Table 1; Supplementary Table 6); analyses adjusted for covariates and calendar time.
Blinding	Physicians could not be blinded to a visible workflow tool; outcomes were extracted automatically from EMR timestamps; expert reviewers were blinded to each other.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☒ Clinical data
☐	☐ Dual use research of concern
☐	☐ Plants

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s) *State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*

Authentication *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See [ICLAC](#) register) *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*

Specimen deposition *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals *For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*

Wild animals	<i>Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.</i>
Reporting on sex	<i>Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.</i>
Field-collected samples	<i>For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.</i>
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	NCT06902675
Study protocol	Registration record at ClinicalTrials.gov; mapping and version history in Supplementary Tables 10-11; full protocol available from the corresponding author.
Data collection	Data is collected from patient medical records (EHR), lab tests, and Imaging Findings.
Outcomes	Pre-specified implementation objectives described in the Methods; full mapping in Supplementary Table 10.

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☐	☐ Public health
☐	☐ National security
☐	☐ Crops and/or livestock
☐	☐ Ecosystems
☐	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☐	☐ Demonstrate how to render a vaccine ineffective
☐	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐ Increase transmissibility of a pathogen
☐	☐ Alter the host range of a pathogen
☐	☐ Enable evasion of diagnostic/detection modalities
☐	☐ Enable the weaponization of a biological agent or toxin
☐	☐ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links May remain private before publication.	For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.
Files in database submission	Provide a list of all files available in the database submission.
Genome browser session (e.g. UCSC)	Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates	Describe the experimental replicates, specifying number, type and replicate agreement.
Sequencing depth	Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.
Antibodies	Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.
Peak calling parameters	Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.
Data quality	Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.
Software	Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.
Instrument	Identify the instrument used for data collection, specifying make and model number.
Software	Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI

☐ Used

☐ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See [Eklund et al. 2016](#))

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involvement in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.