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.

Proof-of-concept validation of the MCIPHER mnemonic framework for three-dimensional tabletop mass casualty incident simulation: a quasi-experimental study

Azza Yousif

aoyousif@dubaihealth.ae

Dubai Health

Abu Omayer

Institute of Learning, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health

Hadeel Farhan

Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health

Naglaa Mohamed Abdelhamied

Dubai Health

Ali AlRahma

Dubai Health

Salma Abdalla Elmisbah

Dubai Health

Dayol Narayanan

Dubai Corporation for Ambulance Services

Radwa Nour

Institute of Learning, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health

Paurnami Prashanth

Dubai Health

Ameline Saouli

Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health

Hossam Hassan Yussef

Dubai Health

Mohamed Alali

Institute of Learning, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health

Ives Hubloue

UZ Brussel University Hospital

Nabil Zary

Institute of Learning, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health

Research Article

Keywords: mass casualty incident, mnemonic, cognitive aid, tabletop exercise, simulation, triage, Cognitive Load Theory, Bayesian analysis, proof-of-concept

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10812236/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Cognitive aids can reduce extraneous cognitive load during mass casualty incident (MCI) management, yet no validated mnemonic framework sequences across all operational phases, from command activation through scene closure, as a single cognitive scaffold. MCIPHER is a five-phase mnemonic (C.O.M.M.A.N.D., S.O.R.T., C.A.R.E., M.O.V.E., C.L.E.A.R.) designed to support novice providers during MCI simulation. This proof-of-concept study evaluated MCIPHER's effect on team performance, individual confidence, knowledge, and attitudes in a standardised three-dimensional (3D) tabletop MCI exercise.

Methods

A quasi-experimental, two-group design was conducted at a university simulation centre. Twenty-five PGY-1 residents were allocated to Group A (control, n = 12) or Group B (MCIPHER, n = 13), each forming one team. Both groups completed a pre-course MCI curriculum; Group B additionally received the MCIPHER framework as a cognitive aid. Teams independently managed a standardised 25-casualty tabletop scenario scored on 20 critical actions. Primary team-level outcomes (Team Performance Score [TPS], Operational Timeliness Index [OTI], Triage Accuracy Score [TAS]) were analysed descriptively with Bayesian Beta-Binomial estimation. Secondary individual-level outcomes (confidence, knowledge, attitudes) were analysed with non-parametric tests and Holm-Bonferroni correction.

Results

Group B demonstrated higher OTI (55.0% vs 30.0%; Bayesian $P(B > A) = 0.942$), TAS (96.0% vs 80.0%; $P(B > A) = 0.951$), and TPS (72.5% vs 65.0%; $P(B > A) = 0.689$). Performance advantages were phase-specific: Group B outperformed on executive phases (Command + 25.0 pp; Closure + 75.0 pp) but scored lower on the procedural Triage phase (− 37.5 pp). Group B attempted 19 of 20 critical actions versus 15 for Group A. A triage accuracy-process dissociation was observed: Group B achieved higher accuracy (96% vs 80%) despite lower triage process scores (62.5% vs 100%). Both groups showed significant within-group confidence gains (Group A: $\Delta = +1.44$, $p < .001$; Group B: $\Delta = +0.74$, $p = .013$), converging to equivalent post-training levels ($p = .891$). MCIPHER was rated favourably (composite 4.27/5).

Conclusions

MCIPHER shows promising effects on operational timeliness and triage accuracy, with gains concentrated in executive decision-making phases consistent with Cognitive Load Theory. The phase-specificity effect and accuracy-process dissociation are novel findings warranting replication. A multi-site randomised trial with multiple teams per condition is needed to confirm these results.

Trial registration

ClinicalTrials.gov, NCT07838441. Registered on 24/09/2026. Retrospectively registered.

Background

Mass casualty incidents (MCIs) impose extraordinary cognitive demands on healthcare providers, requiring simultaneous management of scene assessment, incident command, triage, treatment, transport coordination, and closure, often under time pressure, resource scarcity, and emotional stress [1]. For novice clinicians, the gap between didactic MCI education and operational readiness is well documented: classroom instruction rarely translates into the coordinated decision-making required during a real incident [2, 3]. Simulation-based training can bridge this gap, but the cognitive complexity of MCI scenarios risks overwhelming learners unless the instructional design explicitly manages cognitive load [4, 5].

Cognitive aids (checklists, algorithms, and mnemonics) offer one approach to reducing extraneous cognitive load by externalising critical steps that would otherwise rely on recall under pressure. In the MCI domain, the most widely studied cognitive aids are triage algorithms such as START, SALT, and JumpSTART, for which pooled accuracy estimates range from approximately 55% to 88% depending on the algorithm, training format, and population [6, 7]. However, these tools address only the triage phase. Broader management frameworks such as METHANE, MIMMS, and CSCATTT provide incident reporting structures or training system architectures but have typically been evaluated as components of multi-day courses rather than as isolated cognitive scaffolds [8, 9]. No validated mnemonic framework currently sequences across all operational phases of an MCI, from initial command activation through scene closure, as a single, integrated cognitive scaffold.

Tabletop exercises (TTX) have emerged as a practical and effective modality for MCI training, particularly for cognitive, team-based, and systems-oriented learning goals such as incident command, communication, and triage reasoning [10, 11]. Three-dimensional tabletop formats using physical models and victim cards (such as the Emergo Train System approach) can achieve comparable cognitive learning outcomes to high-fidelity simulation at lower cost and logistical burden [12, 13, 14]. Yet the literature lacks a universally validated performance measurement framework specific to tabletop MCI exercises [15], and most tabletop evaluations are limited to Kirkpatrick Levels 1 and 2 (reaction and learning), with few extending to Level 3 behavioural observation [2].

The MCIPHER mnemonic framework was developed to address these gaps. MCIPHER provides a sequential cognitive scaffold spanning five MCI phases, each encoded as a sub-mnemonic: C.O.M.M.A.N.D. (command and scene management), S.O.R.T. (triage), C.A.R.E. (treatment), M.O.V.E. (transport), and C.L.E.A.R. (closure and debrief). By mapping the full operational sequence onto a retrievable mnemonic structure, MCIPHER aims to reduce extraneous cognitive load for novice providers

during MCI simulation, enabling them to attempt a broader range of critical actions and to maintain situational awareness across phase transitions.

The present study is a proof-of-concept evaluation of MCIPHER in a standardised 3D tabletop MCI exercise. The primary objective was to compare team-level operational performance (Team Performance Score, Operational Timeliness Index, and Triage Accuracy Score) between a group that received the MCIPHER framework as a pre-exercise cognitive aid and a control group that received standard training only. Secondary objectives were to assess individual-level effects on self-reported confidence, knowledge, and attitudes, and to evaluate participants' perceptions of the mnemonic's usefulness. Given the proof-of-concept design (one team per condition), team-level outcomes are reported descriptively with Bayesian probability estimation rather than as confirmatory hypothesis tests.

Methods

Study design and setting

This proof-of-concept study used a quasi-experimental, two-group design to evaluate whether the MCIPHER mnemonic framework improves team performance during a standardised three-dimensional (3D) tabletop mass casualty incident (MCI) exercise. The study was conducted on 18 November 2025 at the Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU) Simulation Centre, Dubai, United Arab Emirates.

The evaluation was structured around the Kirkpatrick model, integrating Level 1 (reaction: debrief experience and perceived usefulness), Level 2 (learning: self-reported confidence, knowledge, and attitudes), and in-domain simulation performance as a proxy for Level 3 (behaviour: operational performance during the tabletop exercise, noting that transfer to real MCI settings was not assessed) [16]. We report this study in accordance with the extensions to the STROBE statement for healthcare simulation research [17].

Participants and recruitment

Twenty-five postgraduate year-1 (PGY-1) residents were recruited from Dubai Health residency programmes across four specialties: emergency medicine, trauma surgery, general surgery, and anaesthesiology. Inclusion criteria were current PGY-1 status in one of the above specialties, ability to provide informed consent, and availability to complete all study components on the simulation day. Exclusion criteria were prior exposure to MCIPHER-guided training or structured MCI tabletop simulations, and inability to complete pre- and post-assessments.

Pre-course curriculum

All participants completed a standardised two-week didactic curriculum ("Mass Casualty Incident: Prehospital Evaluation and Response") before the simulation day, comprising four modules: (1) Introduction to MCI Prehospital Management; (2) Response Activation and Scene Set-Up; (3A) Triage Systems and Decision-Making, (3B) Treatment and Resource Allocation, and (3C) Evacuation and Transport; and (4) Response Conclusion and Recovery. This ensured equivalent foundational knowledge across groups prior to the experimental comparison.

Group allocation and intervention

Participants were allocated to one of two groups:

Group A (Control, n = 12): Proceeded directly to the tabletop exercise after the shared didactic curriculum, without exposure to MCIPHER.

Group B (MCIPHER intervention, n = 13): Received an additional structured pre-brief (Module 5) immediately before the tabletop exercise, introducing MCIPHER as a cognitive roadmap for sequencing and prioritising MCI response. The framework comprises five sequential phases: C.O.M.M.A.N.D. → S.O.R.T. → C.A.R.E. → M.O.V.E. → C.L.E.A.R. [18].

A crossover design was used during the day to ensure all participants ultimately received the educational content; however, only each team's first tabletop performance was used for the between-group comparison to avoid carryover learning effects. Crossover performance data are not reported in this manuscript and will be analysed separately.

Simulation scenario

Both teams responded independently to the same standardised scenario: a bus crash into a football field during a live match, conducted under simulated environmental constraints (cold, wet conditions with rain audio; limited resources including two ambulances and one bus). The active exercise duration was approximately 60 minutes, followed by a structured debrief (~ 20 minutes).

The tabletop environment included a purpose-built 3D map for incident management and zoning, physical 3D assets (ambulances, police vehicles), victim triage cards aligned with the Emergo Train System approach [12], and role-specific documentation tools (incident log, triage log, treatment log, transport log, logistics log).

The scenario comprised 25 casualties distributed by acuity: 4 RED (immediate), 7 YELLOW (delayed), 10 GREEN (minor), and 4 BLACK (deceased). Two casualties were designed to change category at secondary triage (Casualty #9: GREEN to YELLOW; Casualty #18: YELLOW to RED), requiring reassessment capability.

Facilitation and standardisation

A game master controlled exercise progression, delivered timed cues ("injects"), and served as the staging and logistics lead. Facilitators provided prompts only when a team stalled for ≥ 30 seconds after a cue, using neutral phrasing to minimise coaching. Teams assigned functional leadership roles consistent with prehospital MCI response: Incident Commander, Triage Lead, Treatment Lead, and Transport Lead.

Performance framework: critical actions

Team performance was operationalised through 20 predefined critical actions (CAs) spanning five phases: Command (CAs 1–6), Triage (CAs 7–10), Treatment (CAs 11–14), Transport (CAs 15–18), and Closure (CAs 19–20). The 20 CAs were developed by the study team based on established MCI doctrine, including incident command principles, structured incident reporting (METHANE), staged triage and life-saving interventions (SALT), coordinated transport and hospital distribution, and post-incident stand-down with debrief [19].

Each CA was time-stamped and scored using an Action Coding Key (ACK) by trained observers:

- • 1.0: Completed fully, in correct sequence, within the timing window, without facilitator prompting.
- • 0.5: Partial completion, outside timing tolerance, incomplete execution, or completion only after prompting.
- • 0.0: Not observed, incorrect, or out of sequence.

A ± 2-minute timing tolerance was applied around predefined benchmark windows reflecting expected operational targets (e.g., ≤ 5 minutes for command activation, ≤ 10 minutes for ICS and zone establishment, ≤ 25 minutes for triage and treatment progression, ≤ 35 minutes for first red transport, ≤ 60 minutes for closure). Phase weights (Command 30%, Triage 20%, Treatment 20%, Transport 20%, Closure 10%) were assigned a priori, reflecting the relative complexity and number of CAs per phase: Command encompasses six CAs spanning scene assessment through triage readiness, while Closure comprises two CAs. The sensitivity analysis (see below) examines the robustness of between-group patterns under alternative weighting schemes. The full ACK rubric is provided in Additional file 1 (Table S1).

Outcome measures

Primary outcomes (team-level, descriptive)

Three team-level indices were derived from CA performance and triage accuracy:

- • **Team Performance Score (TPS)**: Sum of all 20 CA scores divided by maximum possible score, expressed as a percentage (TPS = Σ CA scores / 20 × 100).
- • **Operational Timeliness Index (OTI)**: Proportion of CAs completed within their benchmark windows, expressed as a percentage (OTI = on-time CAs / 20 × 100).
- • **Triage Accuracy Score (TAS)**: Proportion of casualties correctly categorised compared with the gold standard, expressed as a percentage (TAS = correct triage / 25 × 100).

Because each condition was represented by a single team (n = 1 per group), these outcomes are reported descriptively. Bayesian and bootstrap approaches were used to estimate posterior probabilities of between-group differences (see Statistical analysis).

Secondary outcomes (individual-level, inferential)

Individual-level data (n = 25) supported inferential analysis:

- • **Self-reported confidence**: Nine Likert-scale items (1–5) administered pre- and post-training, covering MCI command, triage, treatment, transport, and communication domains.
- • **Knowledge**: A 30-item binary test administered post-training, covering MCI management concepts aligned with the didactic curriculum.
- • **Attitudes**: Four Likert-scale items (1–5) administered pre- and post-training, assessing perceived importance and motivation for MCI preparedness.
- • **Debrief experience (Kirkpatrick Level 1)**: Seven Likert-scale items (1–5) assessing the perceived quality, realism, and usefulness of the tabletop exercise and debrief.
- • **MCIPHER perception (Group B only)**: Four Likert-scale items (1–5) assessing the perceived usefulness, clarity, and applicability of the MCIPHER framework.

Statistical analysis

All analyses were conducted in Python 3.11 using scipy (v1.11), numpy (v1.26), and pandas (v2.1). The analysis was pre-specified in a registered analytical protocol comprising 15 modules, developed following recommendations from an expert panel review [20].

Inferential analysis (individual-level, n = 25)

Non-parametric tests were used throughout due to ordinal data and small sample sizes. Between-group comparisons used Mann-Whitney U tests; within-group pre-post changes used Wilcoxon signed-rank tests. Fisher's exact tests were used for categorical variables. Effect sizes were calculated as rank-biserial r = |Z| / √N for all tests [21].

For the nine confidence items, Holm-Bonferroni correction was applied to control the family-wise error rate within each group's pre-post comparisons (9 tests per group) [22].

Baseline non-equivalence assessment

Baseline composite confidence differed between groups (Group A: M = 2.69, SD = 0.80; Group B: M = 3.34, SD = 0.52; p = .022, r = .46). This was addressed through a three-pronged analysis: (a) comparison of post-training scores (to test whether the gap persisted), (b) comparison of change scores (Δ = post − pre per participant, to test whether groups improved differently), and (c) partial correlation of post-confidence with group assignment controlling for baseline confidence (to test whether group membership contributed beyond baseline). Additionally, the correlation between pre-training scores and change scores was examined as an indicator of regression to the mean (RTM).

Descriptive analysis with Bayesian estimation (team-level, n = 1 per group)

Because each condition was represented by a single team, conventional between-group hypothesis testing was not applied to team-level outcomes [23, 24]. Instead, we used three complementary approaches:

1. **Descriptive comparison**: Raw TPS, OTI, and TAS values with percentage-point differences. 2. **Bootstrap resampling**: 10,000 bootstrap iterations resampling individual CA scores (with replacement) within each team to estimate sampling distributions of TPS and OTI. The proportion of iterations where Group B exceeded Group A provided an empirical P(B > A). 3. **Bayesian Beta-Binomial analysis**: For each team-level proportion (TPS, OTI, TAS), observed successes and failures were modelled using a Beta-Binomial framework with an uninformative Beta(1,1) prior. Posterior distributions were compared via Monte Carlo simulation (50,000 draws) to estimate P(B > A) and 95% credible intervals [25, 26].

The Beta-Binomial model treats individual CAs as exchangeable Bernoulli trials; we acknowledge that sequential dependencies between CAs within a scenario may partially violate this assumption. However, the convergence between Bayesian posterior estimates and bootstrap resampling results (which do not impose the same exchangeability structure) provides indirect evidence that this violation does not materially alter the conclusions, and the prior sensitivity analysis (three prior specifications) further confirms that posterior estimates are robust to modelling choices. Results from team-level analyses are interpreted as comparative case evidence intended to inform future replicated studies, not as confirmatory treatment effect estimates.

Sensitivity analyses

Phase-level performance was examined under three weighting schemes: equal weighting (each CA worth 1/20), phase-weighted (each phase contributes equally regardless of number of CAs), and clinical-priority weighting (Command and Triage phases weighted 1.5×, Closure weighted 0.5×). Consistency of between-group patterns across weighting schemes was used to assess robustness.

Exploratory analyses

Spearman rank correlations were used to examine associations between MCIPHER perception scores (Group B only) and post-training confidence, and between post-training confidence and debrief

satisfaction. A triage error taxonomy was constructed by classifying each misclassified casualty by error direction (overtriage vs. undertriage) and acuity category, with error rates compared descriptively between groups.

Reporting standards

This study is reported following the CONSORT-SIM extensions for healthcare simulation research [17], adapted for our quasi-experimental (non-randomised) design using relevant STROBE elements for observational studies. Specifically, we apply CONSORT-SIM items for simulation-specific reporting (simulation environment description, facilitation model, outcome assessment) while using the STROBE framework for items related to non-randomised group allocation. A completed checklist mapping each item is provided in Additional file 2 (Table S10).

Results

Participants and baseline characteristics

Twenty-five PGY-1 residents participated (Group A: n = 12; Group B: n = 13). The groups were comparable on age (Group A: M = 25.0, SD = 1.9; Group B: M = 25.7, SD = 1.5; U = 52.5, p = .156, r = .28), gender (male: Group A 33% vs Group B 23%; Fisher's exact p = .673), prior MCI training (Group A 8% vs Group B 8%; p = 1.000), and prior MCI participation (Group A 8% vs Group B 0%; p = .480). Specialty distribution differed descriptively, with a higher proportion of emergency medicine residents in Group B (54% vs 33%), though no formal comparison was planned for this variable. Full demographic data are presented in Table 1.

Baseline non-equivalence assessment

Baseline composite confidence differed significantly between groups (Group A: M = 2.69, SD = 0.80; Group B: M = 3.34, SD = 0.52; U = 35.5, p = .022, r = .46), with Group B reporting higher pre-training confidence. A three-pronged analysis was conducted to evaluate the implications of this non-equivalence (Table 2).

First, post-training composite confidence did not differ between groups (Group A: M = 4.14, SD = 0.61; Group B: M = 4.09, SD = 0.65; U = 81.0, p = .891, r = .03), indicating convergence. Second, change-score comparison showed that while Group A exhibited numerically larger gains (Δ = +1.44, SD = 0.81 vs Δ = +0.74, SD = 0.82), this difference was not significant (U = 111.0, p = .077, r = .36). Third, partial Spearman correlation of post-confidence with group assignment, controlling for baseline confidence, showed no residual group effect (partial ρ = −0.168, p = .433). A negative correlation between baseline confidence and confidence change (ρ = −0.596, p = .002) was consistent with regression to the mean (RTM), suggesting that the larger gains observed in Group A were at least partly attributable to their lower starting point rather than differential training effects. Taken together, these analyses indicate that both groups reached equivalent confidence levels post-training and that no group effect remained after accounting for baseline differences.

Individual-level outcomes (Kirkpatrick Level 2)

Self-reported confidence

Both groups demonstrated significant within-group confidence gains. Group A showed significant improvement on all nine items after Holm-Bonferroni correction (all adjusted $p < .05$; composite $\Delta = +1.44$, $W = 0.0$, $p < .001$, $r = .88$). Group B showed significant composite improvement ($\Delta = +0.74$, $W = 11.0$, $p = .013$, $r = .67$), though only one of nine items survived Holm-Bonferroni correction (conf7: communication and coordination; $\Delta = +1.08$, $W = 0.0$, adjusted $p = .001$, $r = .89$). The remaining eight items in Group B showed consistent positive trends (Δ range: +0.46 to + 0.92) with large uncorrected effect sizes ($r = .44–.73$) but did not survive multiple comparison correction. This pattern reflects Group B's higher baseline scores compressing the available range for measurable improvement (Table 3; Figs. 3 and 7).

Group A showed numerically larger gains than Group B across all nine items (Table 3), with the greatest differences on items assessing triage decision-making (conf3: Δ difference = + 1.23) and scene management (conf2: Δ difference = + 1.15). Despite these item-level differences, the between-group comparison of confidence change scores was not significant at the composite level ($p = .077$), and no individual item-level between-group comparison reached significance after correction.

Knowledge

Post-training knowledge was assessed using a 30-item binary test. Because different quiz versions were administered to the two groups to reflect their differing curricular content, direct between-group comparison was not appropriate and knowledge cannot inform the between-group comparison of MCIPHER’s effect. Within-group descriptive results are reported for completeness: both groups demonstrated strong knowledge acquisition (Group A: M = 89.7%, SD = 4.6; Group B: M = 84.4%, SD = 11.0). Group B exhibited greater variance (Levene's $F = 3.47$, $p = .075$), suggesting more heterogeneous learning outcomes. Item difficulty analysis identified question 19 as the most challenging item for both groups (Group A: 8.3% correct; Group B: 15.4% correct). Full item-level results are presented in Table S2.

Attitudes

Attitude item 1 (self-assessed MCI preparedness) showed significant improvement in both groups (Group A: $\Delta = +2.50$, $p < .001$, $r = .88$; Group B: $\Delta = +1.62$, $p = .001$, $r = .89$), with no significant between-group difference in change scores ($U = 111.0$, $p = .065$, $r = .36$). Attitude items 2–4 (general attitudes toward MCI training importance and motivation) were already high at baseline ($M > 4.4$) and showed no significant pre-post change in either group (all $p > .25$), consistent with a ceiling effect (Table S3).

Kirkpatrick Level 1: Reaction

Both groups rated the debrief experience highly, with no significant between-group differences (composite: Group A M = 4.70, SD = 0.32; Group B M = 4.41, SD = 0.63; $U = 99.0$, $p = .258$, $r = .23$). All

individual debrief items exceeded 4.0 in both groups (Table S4).

Among Group B participants, MCIPHER was rated favourably across all four perception items: overall usefulness (M = 4.23, SD = 0.70), perceived helpfulness (M = 4.46, SD = 0.63), ease of recall (M = 3.92, SD = 0.83), and intent to apply (M = 4.46, SD = 0.63). All items were significantly above the scale midpoint (all $p < .005$ on one-sample Wilcoxon tests against 3.0). The composite MCIPHER perception score was 4.27 (SD = 0.62; 95% bootstrap CI [3.92, 4.60]) (Table S5).

Team-level outcomes (in-domain simulation performance)

Because each condition was represented by a single team, the following results are reported descriptively with Bayesian probability estimates; they do not constitute confirmatory hypothesis tests. P(B > A) values represent the posterior probability that Group B outperformed Group A on a given metric, given the observed data and an uninformative prior; for example, P(B > A) = 0.94 indicates a 94% posterior probability of superiority for Group B.

Overall team performance

Group B achieved a Team Performance Score (TPS) of 72.5% compared with 65.0% for Group A (Δ = +7.5 percentage points). Bayesian Beta-Binomial analysis with an uninformative prior yielded P(TPS_B > TPS_A | data) = 0.689 (Group A posterior: Beta(14.0, 8.0), mean = 63.6%, 95% CrI [43.0, 81.9]; Group B posterior: Beta(15.5, 6.5), mean = 70.5%, 95% CrI [50.3, 87.1]). Bootstrap resampling (10,000 iterations) produced a consistent estimate: P(TPS_B > TPS_A) = 0.703, with a mean difference of 7.4 percentage points (95% CI [− 15.0, 30.0]) (Table 4; Fig. 2).

The Operational Timeliness Index (OTI) showed a larger between-group separation. Group B completed 11 of 20 CAs within benchmark timing windows (55.0%) compared with 6 of 20 for Group A (30.0%; Δ = +25.0 percentage points). Bayesian analysis yielded P(OTI_B > OTI_A | data) = 0.942 (Group A posterior: Beta(7, 15), 95% CrI [14.6, 51.9]; Group B posterior: Beta(12, 10), 95% CrI [34.0, 74.4]). Bootstrap resampling corroborated this finding: P(OTI_B > OTI_A) = 0.931 (Fig. 1; Table S6).

Initial Triage Accuracy Score (TAS) was 96.0% (24/25 correct) for Group B and 80.0% (20/25 correct) for Group A (Δ = +16.0 percentage points). Bayesian analysis yielded P(TAS_B > TAS_A | data) = 0.951 (Group A posterior: Beta(21, 6), 95% CrI [60.6, 91.0]; Group B posterior: Beta(25, 2), 95% CrI [80.4, 99.1]). Secondary triage accuracy converged to 92.0% in both groups (23/25 correct) (Table 4; Fig. 2).

Completeness and independence

Group B attempted 19 of 20 CAs compared with 15 of 20 for Group A, a difference of four additional critical actions. Group A missed five CAs entirely (scored 0.0): CA-1 (declare MCI), CA-3 (send METHANE report), CA-15 (hospital plan confirmed), CA-18 (transport updates sent), and CA-20 (debrief start). Group B missed only one CA (CA-15: hospital plan confirmed). However, Group A demonstrated a higher independence ratio (73.3% vs 52.6%), reflecting that a larger proportion of Group A's attempted CAs were

completed without facilitator prompting (11/15 vs 10/19). This pattern is consistent with the possibility that the MCIPHER scaffold expanded the scope of actions attempted while simultaneously increasing reliance on prompting for some of those additional actions, though this interpretation requires replication with multiple teams to distinguish framework effects from team-specific dynamics (Fig. 4; Table S9).

Phase-level performance and sensitivity analyses

Phase-level analysis revealed a non-uniform pattern of between-group differences (Table 5; Fig. 5). The largest advantages for Group B occurred in the Command phase (75.0% vs 50.0%; Δ = +25.0 percentage points) and Closure phase (75.0% vs 0.0%; Δ = +75.0 percentage points), both phases directly scaffolded by MCIPHER sub-mnemonics (C.O.M.M.A.N.D. and C.L.E.A.R., respectively). Group A scored 0% on Closure, having failed to attempt either closure CA (formal stand-down and debrief). Timing data (Table S9) indicate that Group A's last recorded CA (CA-19: incident stand-down) occurred at 3309 seconds but was scored 0.0, suggesting the team reached the closure window but did not execute structured closure actions; Group B completed CA-19 at 3465 seconds and CA-20 at 3495 seconds.

In contrast, Triage phase scores favoured Group A (100.0% vs 62.5%; Δ = −37.5 percentage points), and Treatment and Transport phases were identical across groups (87.5% and 62.5%, respectively). This differential pattern, in which the group receiving MCIPHER performed better on executive and organisational phases (Command, Closure) but not the procedural Triage phase, is explored further in the Discussion.

Three weighting schemes were applied to assess robustness: equal weighting (each CA weighted 1/20), phase-weighted (each phase contributing equally regardless of CA count), and clinical-priority weighting (Command and Triage phases weighted 1.5×, Closure weighted 0.5×). Group B's overall advantage persisted across all three schemes (equal: Δ = +12.5 pp; original: Δ = +7.5 pp; clinical: Δ = +1.2 pp), though the magnitude varied with the emphasis placed on Command/Closure versus Triage phases (Table 5).

Triage error analysis

Group A committed five initial triage errors (20.0% error rate), all in the direction of overtriage: one BLACK casualty assigned RED (Casualty #2: severe, with clinical signs of non-survivable brain injury), one GREEN assigned YELLOW (Casualty #8: moderate), and three YELLOW casualties assigned RED (Casualties #12, #16, #18: severe). Group B committed one initial triage error (4.0% error rate): one GREEN casualty assigned YELLOW (Casualty #6: moderate overtriage) (Table 6; Fig. 6).

Error distribution by gold-standard category revealed that both groups achieved 100% accuracy for RED (immediate) casualties. Group A's errors were concentrated in the YELLOW category (3/7 correct, 57.1%) and BLACK category (3/4 correct, 75.0%), while Group B achieved 100% accuracy in both. GREEN category accuracy was identical (both 10/11, 90.9%).

A dissociation emerged between triage accuracy and triage process fidelity: Group B achieved higher accuracy (96.0% vs 80.0%) but lower triage-phase CA scores (62.5% vs 100.0%), while Group A demonstrated perfect triage process adherence yet made more categorisation errors. This accuracy-process dissociation suggests that the constructs of "triage correctly" and "follow triage procedures completely" are related but non-interchangeable, and that the MCIPHER group may have prioritised accuracy over procedural documentation, though this observation is based on a single team per condition.

At secondary triage, Group A corrected two initial errors but persisted with one (Casualty #2 remained RED), while introducing one new error (Casualty #9: YELLOW to RED). Group B introduced two new errors at secondary triage (Casualty #1: YELLOW to GREEN; Casualty #18: RED left as YELLOW), both involving the two casualties designed to change category between assessment rounds. Both groups converged to 92.0% secondary triage accuracy.

Exploratory correlations

Among Group B participants, MCIPHER perception scores correlated strongly with post-training confidence ($\rho = .897$, $p < .001$), suggesting that participants who found the mnemonic most useful also reported the highest confidence gains; however, because both measures are self-reported on similar Likert scales, common method variance may inflate this association. Across all participants, post-training confidence was moderately correlated with debrief satisfaction ($\rho = .665$, $p < .001$). Within each group, confidence gain was negatively correlated with post-training knowledge, though neither reached significance (Group A: $\rho = -.516$, $p = .086$; Group B: $\rho = -.543$, $p = .055$). Full correlational data are presented in Table S8.

Table 1
Demographics and baseline characteristics

Characteristic	Group A (n = 12)	Group B (n = 13)	Test statistic	p
Age, M (SD)	25.0 (1.9)	25.7 (1.5)	U = 52.5, r = .28	.156
Male, n (%)	4 (33%)	3 (23%)	Fisher's exact	.673
Specialty				
Emergency Medicine	4 (33%)	7 (54%)		—
Trauma Surgery	3 (25%)	1 (8%)		
Anesthesiology	3 (25%)	3 (23%)		
General Surgery	2 (17%)	2 (15%)		
Prior MCI training, n (%)	1 (8%)	1 (8%)	Fisher's exact	1.000
Prior MCI participation, n (%)	1 (8%)	0 (0%)	Fisher's exact	.480
Baseline confidence, M (SD)	2.69 (0.80)	3.34 (0.52)	U = 35.5, r = .46	.022*

*p < .05. Baseline confidence is the composite of nine self-reported items (1–5 Likert scale). Mann-Whitney U for continuous variables; Fisher's exact for categorical.

Table 2
Baseline non-equivalence assessment (three-pronged analysis)

Analysis	Group A	Group B	Test	Result
Post-training confidence, M (SD)	4.14 (0.61)	4.09 (0.65)	U = 81.0, r = .03	p = .891
Change score Δ, M (SD)	+ 1.44 (0.81)	+ 0.74 (0.82)	U = 111.0, r = .36	p = .077
Partial ρ (post ~ group | baseline)	—	—	Partial Spearman	ρ = −0.168, p = .433
Baseline × change correlation	—	—	Spearman	ρ = −0.596, p = .002*

*p < .05. The three-pronged analysis demonstrates that: (1) both groups converged to equivalent post-training confidence; (2) between-group change-score difference was not significant; and (3) no group effect remained after controlling for baseline. The strong negative baseline×change correlation is consistent with regression to the mean.

Table 3
Self-reported confidence: pre-post within-group results with Holm-Bonferroni correction

Item	A Pre	A Post	A Δ	A p (adj)	B Pre	B Post	B Δ	B p (adj)
conf1: MCI knowledge	2.33	3.83	+ 1.50	.002*	3.08	4.00	+ 0.92	.056
conf2: Scene management	2.17	4.08	+ 1.92	.001*	2.92	3.69	+ 0.77	.114
conf3: Triage decisions	2.00	4.00	+ 2.00	< .001*	3.15	3.92	+ 0.77	.074
conf4: Treatment priorities	2.50	4.00	+ 1.50	.003*	3.31	3.92	+ 0.62	.086
conf5: Resource allocation	2.92	4.25	+ 1.33	.007*	3.69	4.31	+ 0.62	.086
conf6: Transport coordination	3.42	4.33	+ 0.92	.043*	3.77	4.23	+ 0.46	.183
conf7: Communication	3.08	4.25	+ 1.17	.014*	3.31	4.38	+ 1.08	.001*
conf8: Team leadership	2.75	4.08	+ 1.33	.004*	3.31	4.00	+ 0.69	.020
conf9: Overall preparedness	3.08	4.42	+ 1.33	.008*	3.54	4.31	+ 0.77	.018
Composite	**2.69**	**4.14**	**+ 1.44**	**< .001***	**3.34**	**4.09**	**+ 0.74**	**.013***

*Significant after Holm-Bonferroni correction (p < adjusted α). Wilcoxon signed-rank test (W) used for within-group comparisons. Group A: 9/9 items survived correction. Group B: 1/9 items survived correction (conf7 only). Composite = mean of nine items. Item descriptions reflect the confidence domain assessed.

Table 4
Team-level performance outcomes with Bayesian estimation

Metric	Group A	Group B	Bayesian P(B > A)	Bootstrap P(B > A)
TPS	65.0%	72.5%	0.689 (CrI: 43.0–81.9 vs 50.3–87.1)	0.703 (CI: −15.0, 30.0)
OTI	30.0% (6/20)	55.0% (11/20)	0.942 (CrI: 14.6–51.9 vs 34.0–74.4)	0.931
TAS (initial)	80.0% (20/25)	96.0% (24/25)	0.951 (CrI: 60.6–91.0 vs 80.4–99.1)	—
TAS (secondary)	92.0% (23/25)	92.0% (23/25)	—	—
CAs attempted	15/20	19/20	—	—
Independence ratio	73.3% (11/15)	52.6% (10/19)	—	—

TPS = Team Performance Score (weighted average of phase scores). OTI = Operational Timeliness Index (proportion of CAs completed within benchmark ± 2 min). TAS = Triage Accuracy Score. CrI = 95% credible interval (Bayesian). CI = 95% bootstrap confidence interval. Bayesian analysis uses Beta(1,1) uninformative prior. Bootstrap based on 10,000 iterations.

Table 5
Phase-level performance scores and sensitivity analysis

Phase	CAs	A score	B score	Δ (pp)	Equal wt	Phase wt	Clinical wt
Command	1–6	50.0%	75.0%	+ 25.0	—	—	—
Triage	7–10	100.0%	62.5%	−37.5	—	—	—
Treatment	11–14	87.5%	87.5%	0.0	—	—	—
Transport	15–18	62.5%	62.5%	0.0	—	—	—
Closure	19–20	0.0%	75.0%	+ 75.0	—	—	—
Overall TPS	1–20	65.0%	72.5%	+ 7.5	+ 12.5	+ 7.5	+ 1.2

Phase scores = (sum of CA scores in phase / max possible) × 100. Sensitivity analysis weighting schemes: Equal = each CA 1/20; Phase = each phase weighted equally; Clinical = Command and Triage 1.5×, Closure 0.5×. Group B advantage persists across all schemes but magnitude varies with emphasis on executive vs procedural phases.

Table 6
Triage accuracy by casualty category

Category	n	A initial	A correct	B initial	B correct	A 2nd	B 2nd	Notes
RED	4	4/4	100%	4/4	100%	100%	100%	
YELLOW	7	4/7	57.1%	7/7	100%	71.4%	85.7%	A: 3 overtriaged to RED
GREEN	10 + 1*	10/11	90.9%	10/11	90.9%	90.9%	90.9%	Both: 1 error
BLACK	4	3/4	75.0%	4/4	100%	75.0%	100%	A: 1 → RED
Total	25	20/25	80.0%	24/25	96.0%	23/25	23/25	

*One casualty classified GREEN in initial triage with gold-standard change to different category at secondary. All Group A errors were in the direction of overtriage. Group B committed one overtriage error (GREEN→YELLOW). Secondary triage: Group A corrected 2/5 initial errors, introduced 1 new error; Group B introduced 2 new errors involving casualties designed to change category.

Discussion

Principal findings

This proof-of-concept study evaluated the MCIPHER mnemonic framework as a cognitive scaffold for 3D tabletop MCI simulation training among PGY-1 residents. Three headline findings emerged. First, MCIPHER was associated with higher operational timeliness (OTI: 55.0% vs 30.0%; Bayesian P(B > A) = 0.942) and triage accuracy (TAS: 96.0% vs 80.0%; P(B > A) = 0.951), with a more modest advantage in overall team performance (TPS: 72.5% vs 65.0%; P(B > A) = 0.689). Second, these advantages were not uniform across MCI phases: Group B outperformed Group A on executive and organisational phases (Command: +25.0 pp; Closure: +75.0 pp) but scored lower on the procedural Triage phase (− 37.5 pp), revealing a phase-specificity effect. Third, both groups converged to equivalent post-training confidence levels (~ 4.1/5; p = .891) despite a significant baseline gap (p = .022), consistent with a shared training effect modulated by regression to the mean.

Theoretical interpretation

Cognitive Load Theory

The phase-specificity pattern aligns with Cognitive Load Theory (CLT) predictions for cognitive aids in complex task environments [1, 5]. MCIPHER functions as an extraneous load reducer by pre-structuring the mental model for incident management: rather than retrieving the sequence of required actions from

long-term memory under time pressure, the learner follows a mnemonic scaffold that externalises the command structure (C.O.M.M.A.N.D.) and closure protocol (C.L.E.A.R.) [4].

This benefit was concentrated in phases that demand executive decision-making (scene assessment, ICS activation, role assignment, debrief initiation) rather than in the Triage phase, which relies on procedural pattern recognition (matching physiological signs to acuity categories). This distinction between executive and procedural task types has theoretical precedent: Young et al. found that team leader roles carry the highest mental workload in simulated paediatric resuscitation [27], and Andersen et al. demonstrated that CLT-based instructional design for a purely procedural VR task produced no benefit and even negative transfer [28]. The MCIPHER results extend this reasoning to MCI simulation: mnemonics scaffold decision sequencing effectively, but do not substitute for the perceptual pattern recognition required in triage.

Group A's superior Triage phase score (100.0% vs 62.5%) despite lower accuracy (80.0% vs 96.0%) further supports this interpretation. Group A performed the triage CAs correctly (documenting each step of the process) but miscategorised more casualties, while Group B achieved higher accuracy but with less complete procedural documentation. This dissociation suggests that MCIPHER directed cognitive resources toward getting the categorisation right rather than toward thorough documentation of the triage process, representing a reallocation of germane load from process fidelity to outcome accuracy [29]. Specifically, Group B's lower triage-phase CA scores reflect incomplete documentation steps (CAs 9–10: triage log updates and verbal reports to command) rather than failure to perform triage itself. Whether this trade-off is clinically benign (sacrificing optional documentation under time pressure) or concerning (missing essential communication) cannot be resolved from the current data and warrants dedicated investigation with finer-grained process measures.

Zone of Proximal Development and scaffolding

Group B's lower independence ratio (52.6% vs 73.3%) can be interpreted through Vygotsky's Zone of Proximal Development (ZPD) framework. Groot et al. argue that simulation is most productive when learners operate at the frontier of their capability with appropriate support structures [30]. Group B attempted 19 of 20 CAs (compared with 15 for Group A), suggesting that MCIPHER expanded the scope of actions deemed achievable, pushing learners into their ZPD, while simultaneously increasing reliance on facilitator prompting for some of those additional actions. This is consistent with the scaffolding literature in health professions education, where structured supports enable performance beyond unassisted capacity before being gradually faded [31, 32].

The completeness difference (19/20 vs 15/20 CAs) aligns with cognitive aid research showing that structured prompts reduce action omissions. Hall et al. reported that the TEMPIST cognitive aid reduced critical omissions by more than 50% in emergency team performance [33], and our finding extends this to a mnemonic scaffold for MCI command.

Comparison with existing literature

MCI cognitive aids and triage accuracy

Group B's initial triage accuracy of 96.0% and Group A's 80.0% should be interpreted in the context of the tabletop format, where casualties present as written cards with explicit clinical descriptors rather than as ambiguous field presentations. These conditions simplify the perceptual demands of triage relative to field settings, and the known acuity distribution (4 RED, 7 YELLOW, 10 GREEN, 4 BLACK) may have further reduced cognitive load. Published accuracy rates for field-based triage algorithms range widely (55–88%) depending on algorithm, training, and population [6, 34, 35], but direct comparison with our tabletop results is not appropriate given the differences in task demands. More broadly, the tabletop format simplifies not only triage but also team performance outcomes: the structured turn-taking, facilitator prompts after 30 seconds of stalling, and absence of competing physical tasks create a support structure unavailable in real MCI settings, and team-level scores should be interpreted accordingly. Group A's overtriage rate of 20% is consistent with typical simulation-based reports (14–24%) [36, 37], while Group B's 4% rate is unusually low for any training format.

Tabletop simulation

Our 3D tabletop approach is consistent with the growing evidence that TTX is particularly effective for cognitive, team-based, and systems-oriented disaster learning goals [10, 11]. The knowledge levels observed (approximately 85 to 90% correct) are comparable to those reported by Chou and Lin in a 3D tabletop MCI curriculum for senior medical students [13]. Our study extends the TTX literature by incorporating Kirkpatrick Levels 1 and 2 alongside in-domain simulation performance (sometimes termed Level 3) within a single session, whereas most TTX evaluations stop at Level 1 or 2 [2]. We note that our team-level performance outcomes reflect behaviour within the tabletop simulation, not transfer to real MCI settings; evidence for transfer would require field-based observation or longitudinal follow-up.

Confidence and metacognitive calibration

The convergence pattern, where a significant pre-training confidence gap ($p = .022$) was eliminated post-training ($p = .891$), is consistent with the broader confidence-calibration literature. Gottlieb et al. have highlighted that simulation-based training compresses self-reported confidence toward the upper end of Likert scales, often with weak correspondence to objective performance [38]. Kerry et al. reported a similar convergence of team efficacy scores across repeated simulation sessions [39]. Our three-pronged analysis (post-score equivalence, change-score comparison, partial correlation controlling for baseline) provides a more rigorous treatment of baseline non-equivalence than most confidence studies, which typically report only raw pre-post comparisons [40].

Novel contributions

This study offers several contributions that, while preliminary given the proof-of-concept design, address identified gaps in the literature.

First, the phase-specificity effect, where MCIPHER enhanced executive phases (Command + 25.0 pp, Closure + 75.0 pp) but not the procedural Triage phase (− 37.5 pp), has not been previously reported. Prior evaluations of MCI mnemonics and cognitive aids have treated performance as a global outcome; our phase-level analysis reveals that the benefit is structurally non-uniform and follows a theoretically predictable pattern.

Second, the triage accuracy-process dissociation (Group B: 96.0% accuracy / 62.5% triage CA score; Group A: 80.0% / 100.0%) provides quantitative evidence that accuracy and process fidelity are related but non-interchangeable constructs. Ingrassia et al. observed a similar pattern (training improved command and control but not triage accuracy) without explicitly framing it as a dissociation [41]. Our data suggest that MCIPHER may redirect cognitive resources from procedural documentation toward categorisation accuracy.

Third, the completeness hypothesis (19/20 vs 15/20 CAs attempted) suggests that MCIPHER's value may lie less in improving the quality of individual actions and more in expanding the range of actions attempted, a structural benefit distinct from accuracy gains.

Fourth, the Bayesian Beta-Binomial approach to single-team comparison addresses a methodological gap identified by Varnell et al. [23] and Murray [24]. With one team per condition, conventional inferential statistics for team-level outcomes are invalid. Our study demonstrates how uninformative priors, posterior probability estimation, and prior sensitivity analysis can provide interpretable evidence under these constraints, and offers sample reporting language for other simulation researchers facing the same design limitation [25].

Limitations

Several limitations constrain the interpretation of these findings. The most important is that each experimental condition was represented by a single team, precluding inferential comparison of team-level outcomes. The Bayesian and bootstrap estimates reported here quantify the probability that one team outperformed the other, not the probability that the MCIPHER framework would be superior in general. Replication with multiple teams per condition is essential before any causal claim can be supported.

The quasi-experimental design without randomisation introduces the possibility of unmeasured confounders. The baseline confidence difference (Group B higher, $p = .022$) and the higher proportion of emergency medicine residents in Group B (54% vs 33%) are examples; although neither demonstrably influenced team-level outcomes, they cannot be fully ruled out. The three-pronged non-equivalence analysis mitigates but does not eliminate this concern.

Beyond these primary limitations, several others merit attention. We conducted the study at a single site with one scenario, which limits generalisability to other institutions and clinical contexts. Scoring was performed by a single investigator using the ACK rubric, and we did not formally assess inter-rater

reliability. The post hoc scoring from video recordings, rather than live observation with blinded observers, introduces potential bias. We collected no retention data beyond the training session itself, so we cannot determine whether confidence and performance gains persist over time. Eye-tracking equipment was present but the data remain unprocessed; these physiological measures would have provided objective indices of cognitive load that self-report cannot capture. The 20-critical-action rubric was developed by the study team based on established MCI doctrine rather than independently validated through expert consensus or item analysis, and future work should incorporate formal content validation. The knowledge assessment employed different quiz versions for the two groups to reflect their differing curricular content, a design choice that precluded direct between-group comparison. We planned qualitative analysis of the debrief transcripts to understand how participants actually used (or did not use) the MCIPHER mnemonic during the exercise, but this analysis was not completed, limiting our insight into the mechanism. Finally, self-reported confidence is recognised as a weak proxy for competence, and ceiling effects at the upper end of the Likert scale (post-training means approximately 4.1/5) may have compressed measurable group differences [38, 40]. Similar ceiling effects were observed in the debrief satisfaction items (all means above 4.0) and attitude items 2 to 4 (baseline means above 4.4), limiting the sensitivity of these instruments to detect between-group differences. Future studies should consider instruments with greater dynamic range or visual analogue scales to mitigate this compression.

Implications

For practice

Both groups achieved high satisfaction ratings (> 4.4/5) and substantial confidence and attitude gains, confirming that the 3D tabletop format with structured debrief is a viable and well-received approach for MCI training among junior residents regardless of mnemonic exposure. MCIPHER itself was rated favourably by Group B (composite 4.27/5; 95% bootstrap CI [3.92, 4.60]) and appears to function as intended, serving as a sequential scaffold for the decision-making demands of MCI command. Educators designing MCI tabletop curricula may consider MCIPHER as a pre-briefing cognitive aid, with the caveat that its effects on triage procedural fidelity require further investigation.

For future research

A definitive evaluation of MCIPHER will require a multi-site cluster-randomised trial with multiple teams per condition, enabling conventional group-level inference. Future studies should include: processed eye-tracking data as a physiological measure of cognitive load; retention assessments at 3- and 6-month intervals; formal inter-rater reliability for the ACK rubric, ideally with at least two independent scorers rating a subset of CAs from video recordings to establish agreement; qualitative debrief analysis to explore how participants used (or did not use) the mnemonic in real time; and scenarios of varying complexity to test the robustness of the phase-specificity finding. The dissociation between triage accuracy and process fidelity warrants dedicated investigation with finer-grained process measures.

Conclusions

This proof-of-concept quasi-experimental study provides preliminary evidence that the MCIPHER mnemonic framework enhances operational timeliness (Bayesian $P(B > A) = 0.942$) and initial triage accuracy ($P(B > A) = 0.951$) in 3D tabletop MCI simulation training among PGY-1 residents. Gains were concentrated in executive decision-making phases (Command, Closure) rather than the procedural Triage phase, suggesting that mnemonic scaffolding is differentially effective across task types. This finding is consistent with Cognitive Load Theory predictions but has not been previously demonstrated in MCI training.

Three novel observations merit replication: the phase-specificity effect, the dissociation between triage accuracy and process fidelity, and the completeness hypothesis (19/20 vs 15/20 critical actions attempted). Both groups demonstrated significant confidence and attitude gains, confirming the educational value of the 3D tabletop format regardless of mnemonic exposure.

Methodologically, this study demonstrates the application of Bayesian Beta-Binomial estimation for the common simulation research scenario of one team per condition, offering a transparent analytic framework where conventional group-level inference is unavailable.

A multi-site cluster-randomized trial with multiple teams per condition is the essential next step to confirm these promising initial findings and establish whether the phase-specificity effect is robust across different scenarios and learner populations.

Declarations

Ethics approval and consent to participate

This study was approved by the Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU) Institutional Review Board [MBRU IRB-2025-717]. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki, except that trial registration was completed retrospectively. All participants provided written informed consent prior to participation. The study was retrospectively registered with ClinicalTrials.gov (NCT07838441).

Consent for publication

Not applicable. No individual participant data are presented.

Availability of data and materials

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request. The MCIPHER framework materials and Action Coding Key rubric are provided in Additional file 1.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the Mohammed Bin Rashid University of Medicine and Health Sciences through the Dubai Health Collaborative Stimulus Research Grant (CSRG) for the M-CIPHER project. The funding organization had no influence on the study design, data analysis, or manuscript preparation.

Authors' contributions

A.Y.: Conceptualization, Methodology, Investigation, Project administration, Supervision, Resources, Writing – original draft, and Writing – review and editing. A.O.: Conceptualization, Methodology, Data curation, Formal analysis, Software, Visualization, Writing – original draft, and Writing – review and editing. H.F.: Conceptualization, Methodology, Investigation, Data curation, Project administration, Resources, Writing – original draft, and Writing – review and editing. N.M.A.: Investigation, Resources, Validation, and Writing – review and editing. M.A.: Investigation, Data curation, Resources, and Writing – review and editing. A.A.: Methodology, Investigation, Validation, and Writing – review and editing. S.A.E.: Investigation, Data curation, Validation, and Writing – review and editing. D.N.: Investigation, Resources, Validation, and Writing – review and editing. R.N.: Investigation, Data curation, and Writing – review and editing. P.P.: Investigation, Project administration, Resources, and Writing – review and editing. A.S.: Investigation, Validation, and Writing – review and editing. H.H.Y.: Investigation, Resources, and Writing – review and editing. I.H.: Methodology, Validation, Supervision, and Writing – review and editing. N.Z.: Conceptualization, Methodology, Funding acquisition, Resources, Project administration, Supervision, and Writing – review and editing.

Acknowledgements

The authors thank Dr. Sarah Kazim, Chair of Emergency Medicine at Dubai Health, for her support and for fostering the departmental environment that helped make this work possible. They also thank Mohammed Bin Rashid University of Medicine and Health Sciences for its support and collaboration in this research, Shakeel Tegginmani of the Al Maktoum Medical Library, for assistance with the literature search, and the Institute of Learning (IOL), for research support.

References

1. Fraser KL, Ayres P, Sweller J. Cognitive load theory for the design of medical simulations. Simul Healthc. 2015;10(5):295–307.
2. Abdelhamied NM, Omayer A, Elmisbah SA et al. Curriculum design in disaster and mass casualty education: a scoping review of healthcare education. BMC Med Educ (2026).
3. Baetzner AS, Schläppi M, Schmid N, Amsler F. Preparing medical first responders for crises: a systematic literature review of disaster training programs. Scand J Trauma Resusc Emerg Med. 2022;30(1):56.

4. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ. 2010;44(1):85–93.
5. Reedy GB. Using cognitive load theory to inform simulation design and practice. Clin Simul Nurs. 2015;11(8):355–60.
6. Franc JM, Kirkland SW, Wisnesky UD, Campbell S, Rowe BH. METASTART: a systematic review and meta-analysis of the diagnostic accuracy of START. Prehosp Disaster Med. 2022;37(1):106–16.
7. Wisnesky UD, Kirkland SW, Rowe BH, Campbell S, Franc JM. A qualitative assessment of studies evaluating the classification accuracy of START: a scoping review. Front Public Health. 2022;10:676704.
8. Hodgetts TJ, Mackway-Jones K. Major Incident Medical Management and Support: The Practical Approach at the Scene. 2nd ed. London: BMJ Publishing Group; 2003.
9. Nunan T, Daud M, Gormley S, Hughes J, Mangan JJ. Review of the national major emergency framework. Int J Disaster Risk Reduct. 2020;51:101811.
10. Frégeau A, Cournoyer A, Bhatt M, et al. Tabletop simulation in health professions education: a scoping review. Simul Healthc. 2024;19(4):234–44.
11. Emaliyawati E, Utami SR, Firmansyah Y, et al. Enhancing disaster preparedness through tabletop disaster exercises: a scoping review. Adv Med Educ Pract. 2025;16:95–107.
12. Lennquist S. The Emergotrain System for training and testing disaster preparedness. Int J Disaster Med. 2003;1(1):29–35.
13. Chou WC, Lin YC. Teaching senior medical students MCI management by 3D tabletop exercise without lecture. BMC Med Educ. 2025;25:846.
14. Rashed EA, Lotfy A et al. Comparing tabletop and high-fidelity simulation for disaster medicine training. Disaster Med Public Health Prep. 2025. Online ahead of print.
15. Skryabina EA, Reedy GB, Amlôt R, Riley P. Developing a performance assessment framework for mass casualty exercises. Int J Disaster Risk Reduct. 2018;31:258–67.
16. Kirkpatrick DL, Kirkpatrick JD. Evaluating Training Programs: The Four Levels. 3rd ed. San Francisco: Berrett-Koehler; 2006.
17. Cheng A, Kessler D, Mackinnon R, et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. Adv Simul. 2016;1:23.
18. Alqahtani N, Zary N. The MCIPHER mnemonic framework for mass casualty incident management: development and rationale. Unpublished doctoral thesis chapter, Mohammed Bin Rashid University of Medicine and Health Sciences; 2025.
19. Advanced Life Support Group (ALSG). Major Incident Medical Management and Support. London: Wiley-Blackwell; 2017.
20. [Analytical protocol available from corresponding author on request.].
21. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen. 2012;141(1):2–18.

22. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70.
23. Varnell SP, Murray DM, Baker WL. An evaluation of analysis options for the one-group-per-condition design. Eval Rev. 2001;25(4):440–53.
24. Murray DM. Design and Analysis of Group-Randomized Trials. New York: Oxford University Press; 1998.
25. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med. 2000;19(24):3417–32.
26. Lui KJ, Mayer JA, Eckhardt L. Confidence intervals for the risk ratio under cluster sampling based on the beta-binomial model. Stat Med. 2000;19(21):2933–42.
27. Young JQ, Irby DM, Cooke M, ten Cate O, Paradis E. Workload of learners during simulated paediatric CPR. BMJ Simul Technol Enhanc Learn. 2021;7:e000640.
28. Andersen SAW, Mikkelsen PT, Konge L, et al. Implementing CLT-based design principles in VR simulation training: a randomized controlled trial. Adv Simul. 2016;1:20.
29. Ambardekar AP, Rosero EB, Bhoja R, et al. Learners' decision-making and task load during a simulated airway crisis. Simul Healthc. 2019;14(2):96–103.
30. Groot F, Kehoe C, Browne J, et al. Simulation at the frontier of the zone of proximal development. Acad Med. 2020;95(7):1098–105.
31. Masava B, Nyoni CN, Botma Y. Scaffolding in health sciences education programmes: an integrative review. Med Sci Educ. 2022;32:791–802.
32. Parker BC, Myrick F. Transformative learning as a context for human patient simulation. J Nurs Educ. 2009;48(6):326–32.
33. Hall AB, Robertson D, Rolfe M, et al. TEMPIST: do cognitive aids reduce error rates in resuscitation? Hum Resour Health. 2020;18:10.
34. Risavi BL, Salen PN, Heller MB, Arcona S. A two-hour intervention using START improves prehospital triage. Prehosp Emerg Care. 2001;5(2):197–9.
35. Chen JY, Shih CL. A study of mass casualty triage simulation training for prehospital EMS providers. Prehosp Disaster Med. 2003;18(1):35–40.
36. Hoff WS, Schwab CW. Trauma system development in North America. Clin Orthop Relat Res. 2004; (422):17–22.
37. Carenzo L, Gamberini L, Crimaldi F, et al. Factors affecting prehospital triage accuracy in simulated MCIs. Scand J Trauma Resusc Emerg Med. 2024;32:87.
38. Gottlieb M, Chan TM, Zaver F, Ellaway RH. Confidence-competence alignment in medical education: a conceptual review. Med Educ. 2022;56(1):15–25.
39. Kerry MJ, Ander DS, Davis JA. Interprofessional simulation's impact on team efficacy beliefs over time. BMJ Simul Technol Enhanc Learn. 2019;6(3):140–7.
40. Sears K, Goldsworthy S, Goodman WM. The relationship between simulation in nursing education and medication safety. J Nurs Educ. 2010;49(1):52–5.

41. Ingrassia PL, Ragazzoni L, Tengattini M, et al. Nationwide program for medical students on disaster medicine in Italy. Prehosp Disaster Med. 2014;29(1):1–7.

Figures

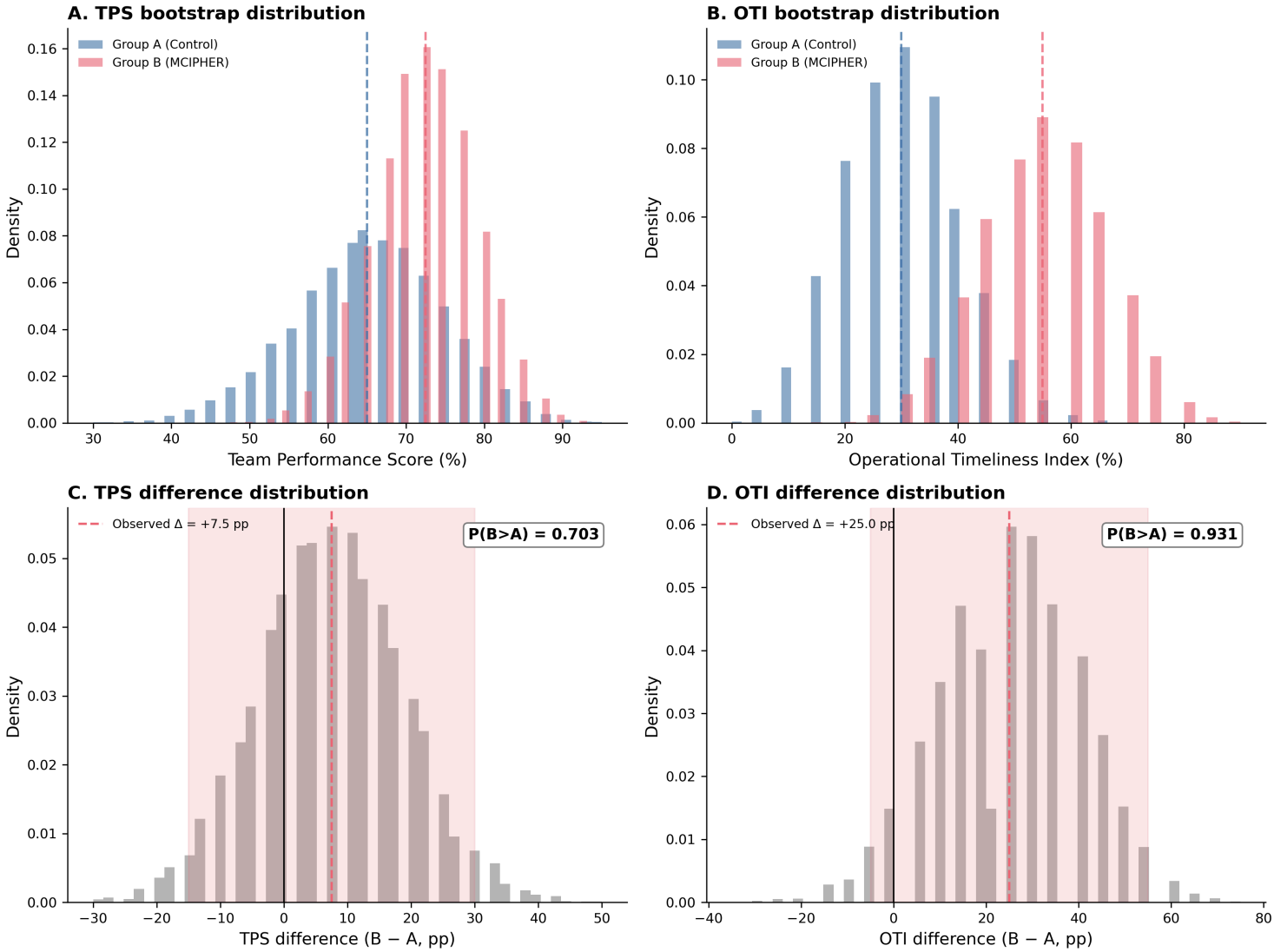


Figure 1

Bootstrap distributions for Team Performance Score (TPS) and Operational Timeliness Index (OTI). Panels A–B show the bootstrap sampling distributions (10,000 iterations) for each metric by group. Panels C–D show the bootstrap difference distributions (Group B − Group A). Dashed vertical lines indicate observed values; shaded regions represent the 95% confidence interval.

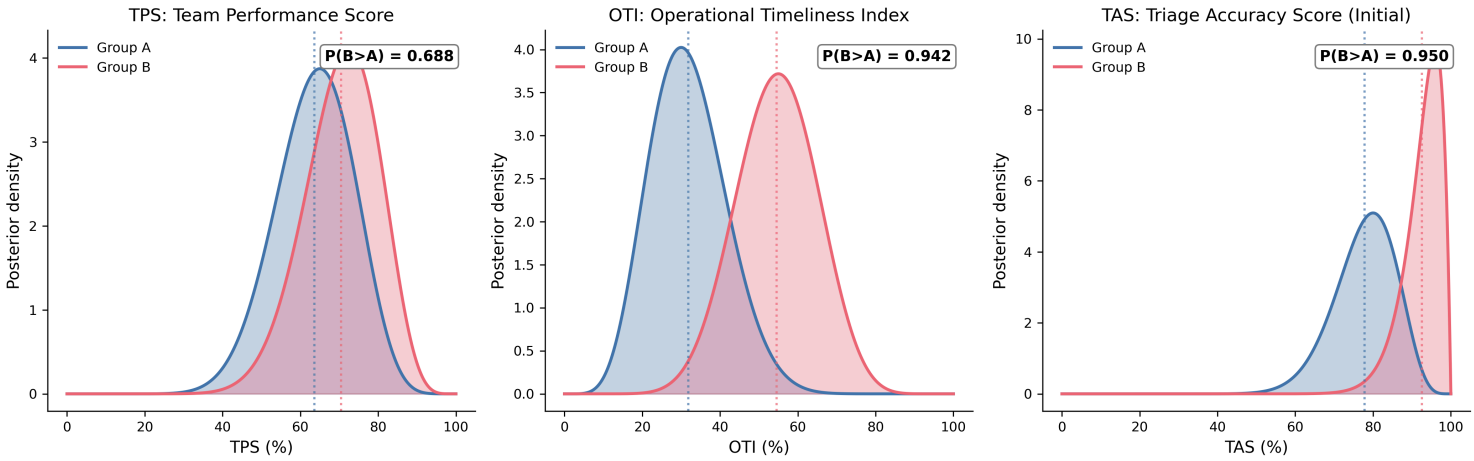


Figure 2

Bayesian posterior probability densities for team-level outcomes. Beta-Binomial posteriors with uninformative Beta(1,1) priors for (A) Team Performance Score, (B) Operational Timeliness Index, and (C) Triage Accuracy Score. Shaded areas represent 95% credible intervals. P(B > A) values indicate the posterior probability that Group B outperformed Group A.

Figure 3. Confidence items: pre–post scores and change by group

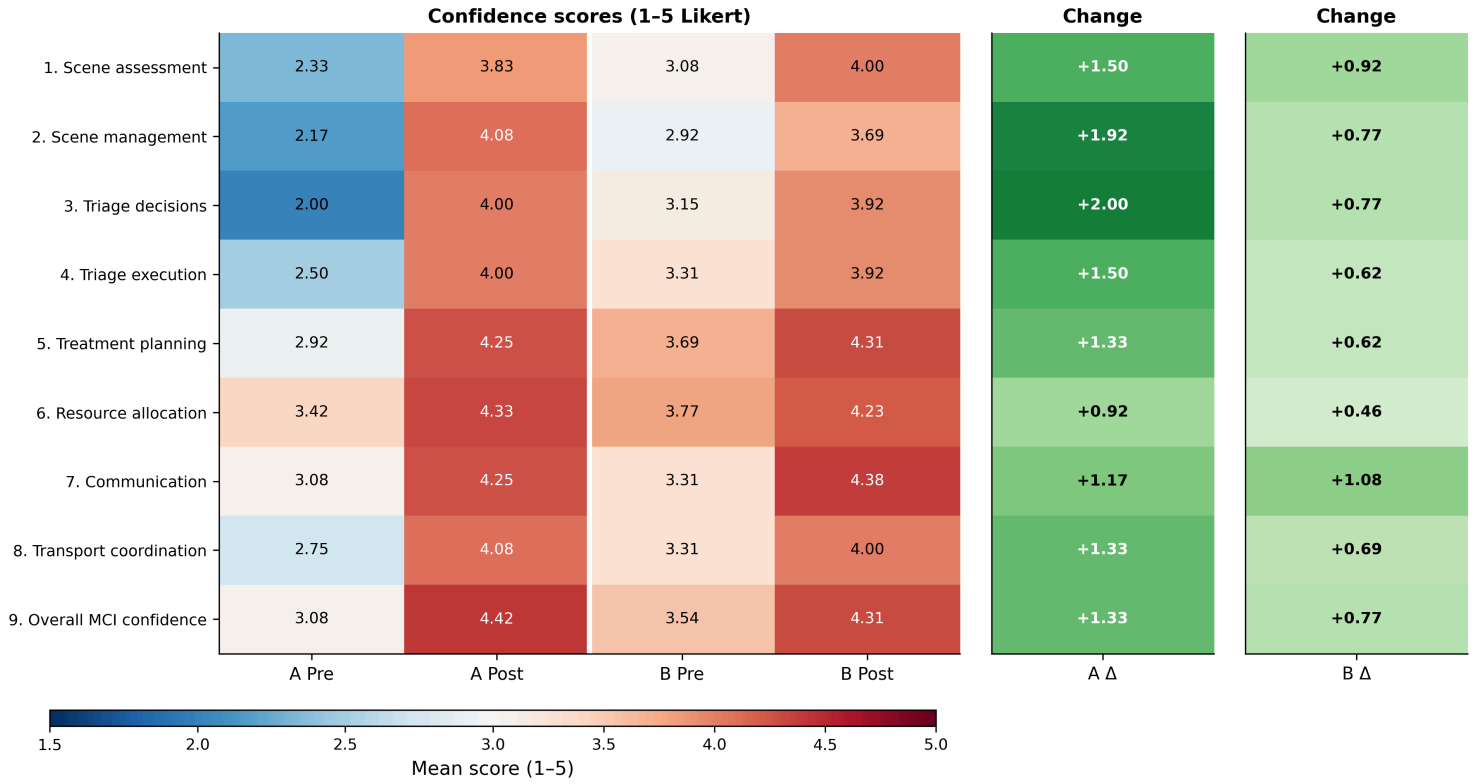


Figure 3

Self-reported confidence heatmap. Mean Likert scores (1–5) for nine confidence items assessed pre- and post-training, by group. Colour scale ranges from 1.0 (low, cool) to 5.0 (high, warm). Δ columns

show within-group change scores.

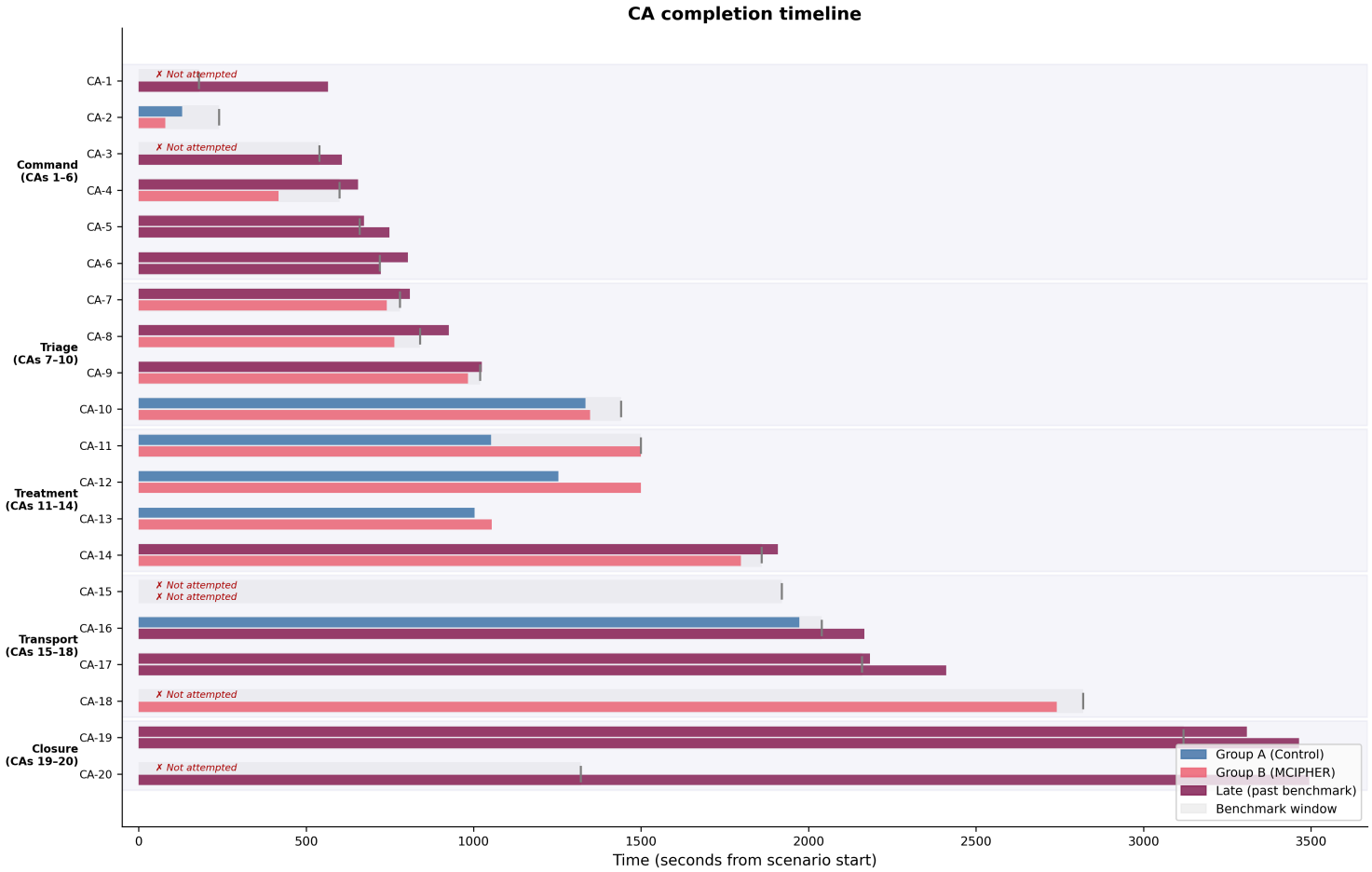


Figure 4

Critical action timeline. Horizontal bars indicate the timing of each critical action (CA) relative to exercise start. Green = on-time; orange = late; deep magenta = substantially late; ✗ = not attempted.

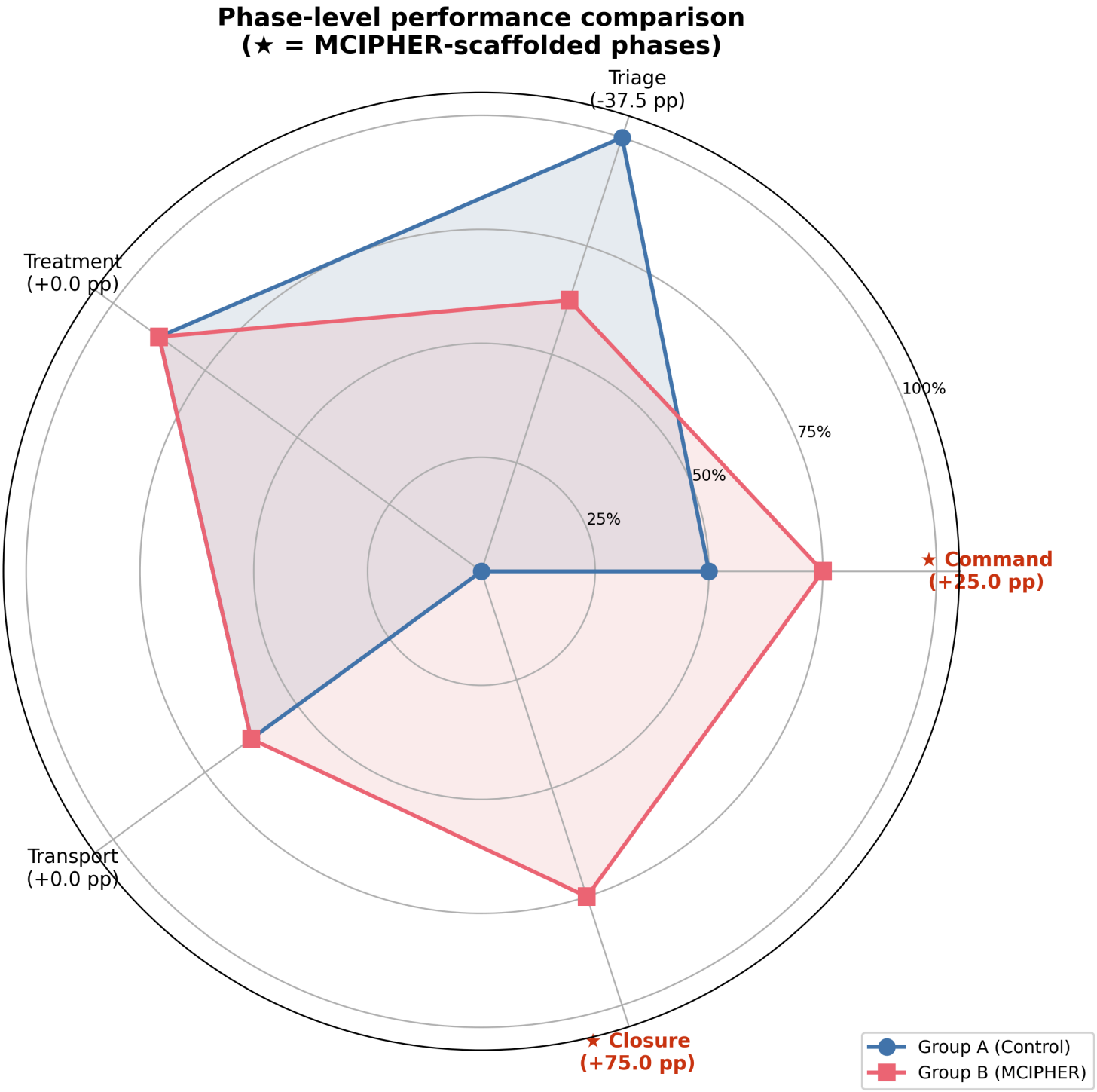


Figure 5

Phase-level performance radar chart. Percentage scores for each of the five MCI phases by group. Phases directly scaffolded by MCIPHER sub-mnemonics are marked with ★.

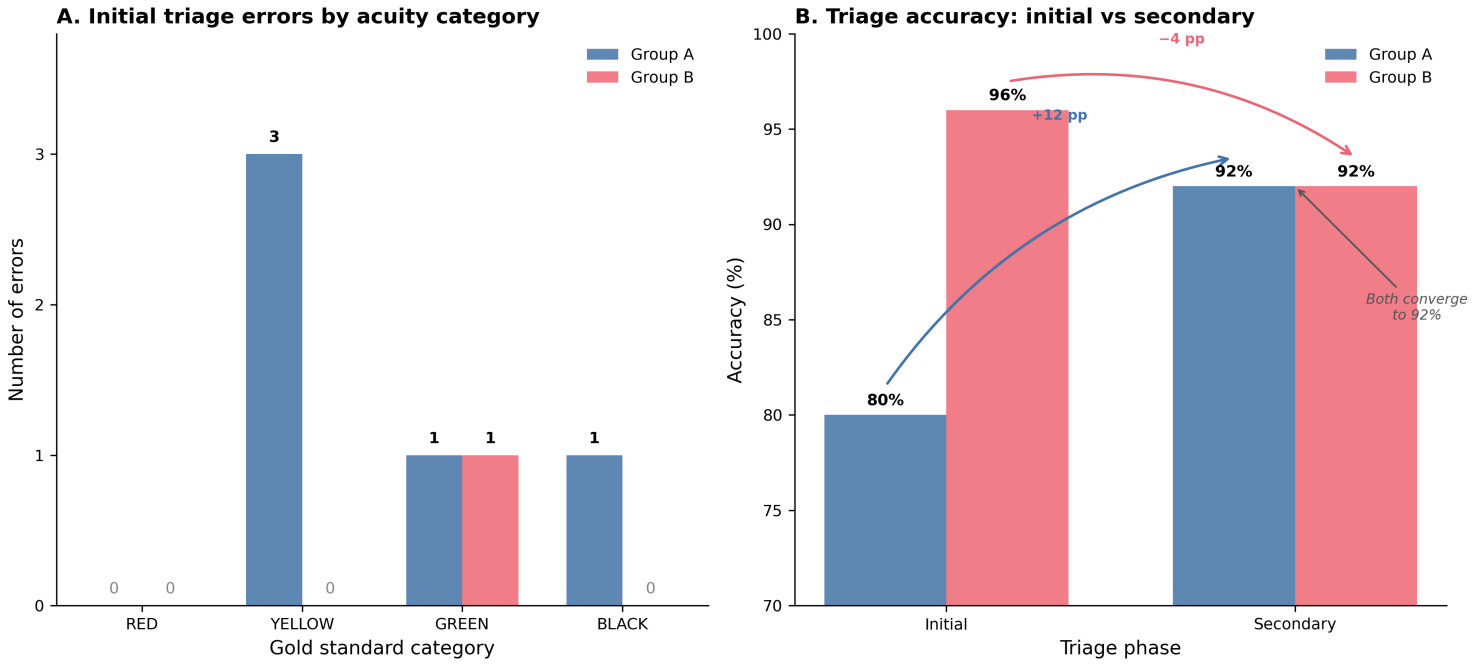


Figure 6

Triage error taxonomy. (A) Initial triage accuracy by gold-standard acuity category. (B) Error counts by direction for initial and secondary triage, with connecting arcs showing error evolution between assessment rounds.

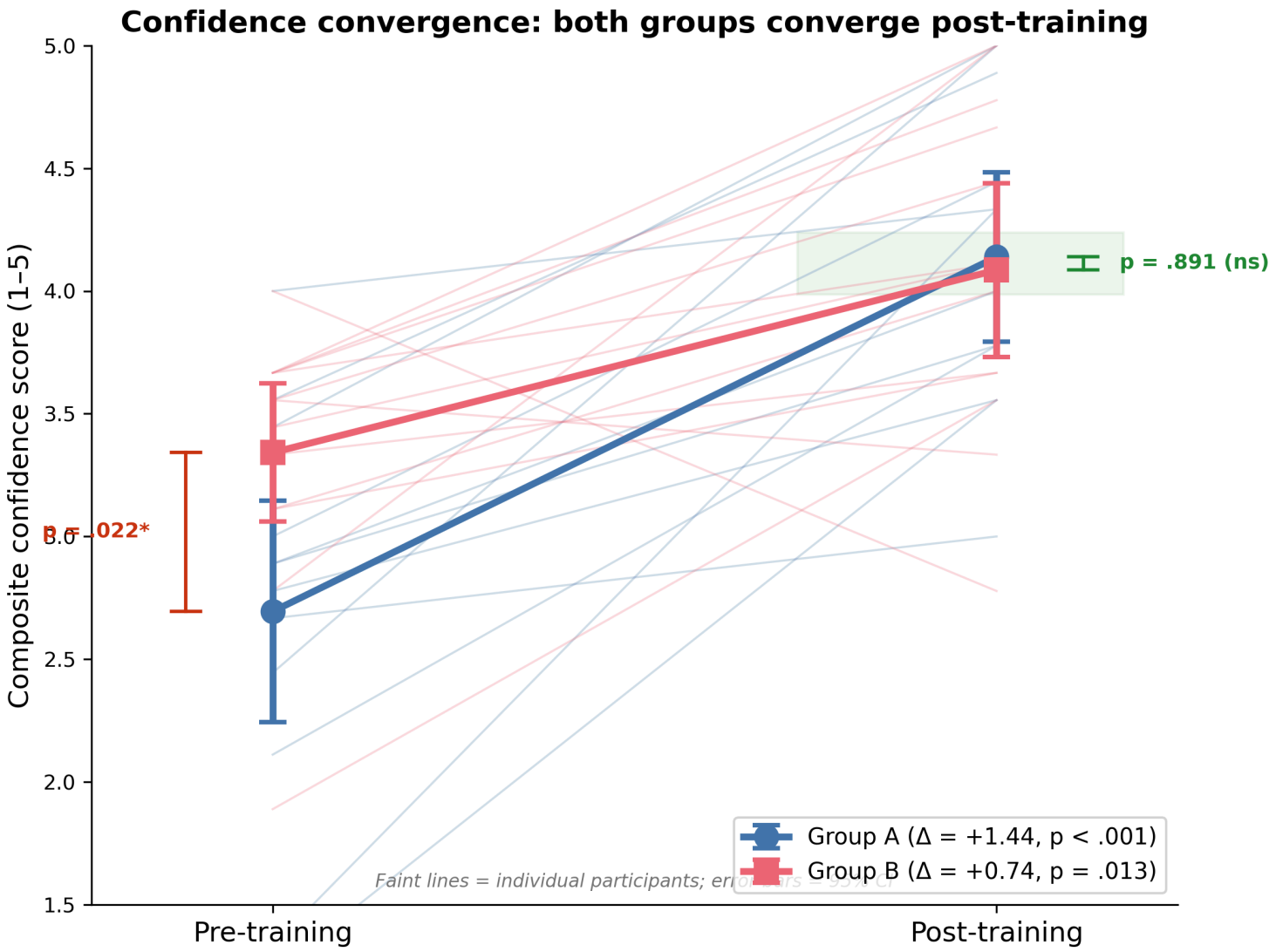


Figure 7

Confidence convergence trajectories. Individual participant pre- to post-training confidence composite scores, by group. Group means connected by thick lines.

Supplementary Files

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

- Additionalfile1SupplementaryMaterials.docx
- Additionalfileslegends.docx