Beyond the Interface: Parameter-Level Documentation Accuracy Across Documentation Methods in Simulated Prehospital Trauma Care

Shani Sultani, MD[1,2], Daniel Gelman, MD[2,3], Danielle Akler, MD[2], and Liora Shmueli, PhD[1]

[1]Department of Management, Bar-Ilan University, Ramat Gan, Israel · [2]Israel Defense Forces Medical Corps, Israel · [3]Department of Military Medicine, Faculty of Medicine, Hebrew University, Jerusalem, Israel.

Abstract

Background: Accurate prehospital trauma documentation is persistently suboptimal. Documentation forms incorporate structurally heterogeneous field types that impose different cognitive demands and may respond differentially to the documentation modality used. This study compared documentation accuracy with paper-based, tablet-based, and artificial intelligence-based speech recognition (AI-SR) methods across ten data fields. It examined whether parameter-level documentation accuracy was consistent across methods.

Methods: A randomized crossover simulation study was conducted among 181 advanced life support (ALS) providers (91 physicians, 90 paramedics) in 36 teams. Each team completed three standardized trauma simulations using all three documentation methods in an order randomized by Latin square. Field-level accuracy was scored against observer-generated reference records.

Results: Parameter rankings were highly concordant across methods (Kendall's W = 0.87), forming three tiers that were identical under all methods: Safety, Evacuation, and Cardiovascular highest (75.6–76.3% overall), Breathing and Airway intermediate (55.4–56.4%), and Resuscitation Round, Personal Data, Secondary Round, Disability, and Injury lowest (37.2–38.1%). AI-SR achieved the highest accuracy for every parameter, followed by tablet and paper (all $P < 0.001$, including Bonferroni-corrected pairwise comparisons); its advantage over paper ranged from 28.6 to 39.2 percentage points.

Conclusions: In this simulation study, documentation accuracy followed a parameter-level hierarchy that was similar across paper, tablet, and AI-SR documentation. AI-SR improved accuracy for every parameter while largely preserving the three-tier structure, suggesting that content-level barriers to documentation persist across interfaces and may require targeted interventions beyond interface design. These findings require confirmation in operational settings.

Keywords: prehospital documentation; speech recognition; artificial intelligence; Tactical Combat Casualty Care; trauma

Corresponding Author: Shani Sultani, MD. Email: Shani.sultani@gmail.com

To cite this article: Sultani S, Gelman D, Akler D, Shmueli L. Beyond the Interface: Parameter-Level Documentation Accuracy Across Documentation Methods in Simulated Prehospital Trauma Care. J Isr Mil Med. 2026;22(69):28-36.

Submitted for publication: 13.12.2025
Approved for publication: 16.2.2026

INTRODUCTION

Accurate and timely documentation of prehospital trauma care is a foundational element of effective trauma systems. Comprehensive prehospital records inform emergency department triage decisions, support continuity of care, and provide the data infrastructure for quality improvement and research [1]. Despite this importance, prehospital documentation remains incomplete. In the IDF Trauma Registry, even after a decade of improvement efforts, 41.2% of casualties in 2018 had no documented prehospital vital sign, and documentation of individual vital signs ranged from 19.1% (pain) to 56.3% (heart rate) [2]. Information gaps also affect civilian care: physicians reported missing previously collected information in almost one-third of emergency department visits, and these gaps were associated with longer stays [3]. In civilian trauma patients, failure of EMS personnel to document basic scene physiology was associated with approximately twice the odds of in-hospital death [4]. Systematically collected prehospital data are what allow trauma systems to be evaluated: a ten-year trauma audit in Iraq used them to show that prehospital trauma care reduced mortality [5]. The paucity of such data limits the development of prehospital trauma research [6].

The cognitive and operational demands of prehospital environments are well recognized as key drivers of documentation failure. Providers must simultaneously manage patient care, team coordination, and environmental stressors, conditions in which manual documentation directly competes with clinical attention [7,8]. This challenge is compounded by the structural heterogeneity of prehospital documentation forms, which incorporate multiple distinct field types. For example, the Tactical Combat Casualty Care (TCCC) card includes free-text entries for patient identification and clinical description, numeric fields for quantitative physiological and pharmacological data, and binary checkbox indicators for performed interventions and assessments. These field types impose qualitatively different cognitive demands.

Existing work on trauma documentation has largely focused on overall form completion rates or method-level comparisons, without examining whether individual parameters show stable documentation patterns across methods [9,10]. An important unanswered question is whether documentation accuracy depends more on the documentation method (the interface used) or on properties intrinsic to the parameter itself, such as its clinical salience, observational accessibility, or the cognitive effort required to retrieve it. If parameter-level documentation accuracy is consistent across methods, this would suggest that interface-level interventions alone are insufficient to address the full scope of the documentation problem.

Three documentation modalities are currently in use or planned for use in prehospital and military medicine: paper-based forms, electronic tablet-based systems, and artificial intelligence-based speech recognition (AI-SR) agents. AI-SR systems use large language models and natural language processing to enable hands-free, real-time documentation from continuous speech [11,12] and have shown promise in hospital settings [13,14]. Evaluation in prehospital trauma care has been limited [15,16], and we are not aware of any study that systematically examined whether the documentation hierarchy across parameters is preserved or disrupted by method.

This study addresses this gap by comparing paper, tablet, and AI-SR documentation accuracy across ten TCCC parameters, primarily assessing whether the relative ranking of parameters by documentation accuracy is consistent across methods, and secondarily examining the magnitude of method-level differences within each parameter.

METHODS

Study Design

This randomized crossover simulation study was conducted at a military medical training facility between April and November 2025. The three documentation methods were administered in sequences defined by a 3 × 3 Latin square (Figure 1); teams were randomly allocated by lots prepared and drawn by a study staff member. The Institutional Review Board reviewed the study and granted exemption from full committee review. Participation was voluntary. All participants provided written

informed consent. Reporting follows the CONSORT extension for randomized crossover trials.

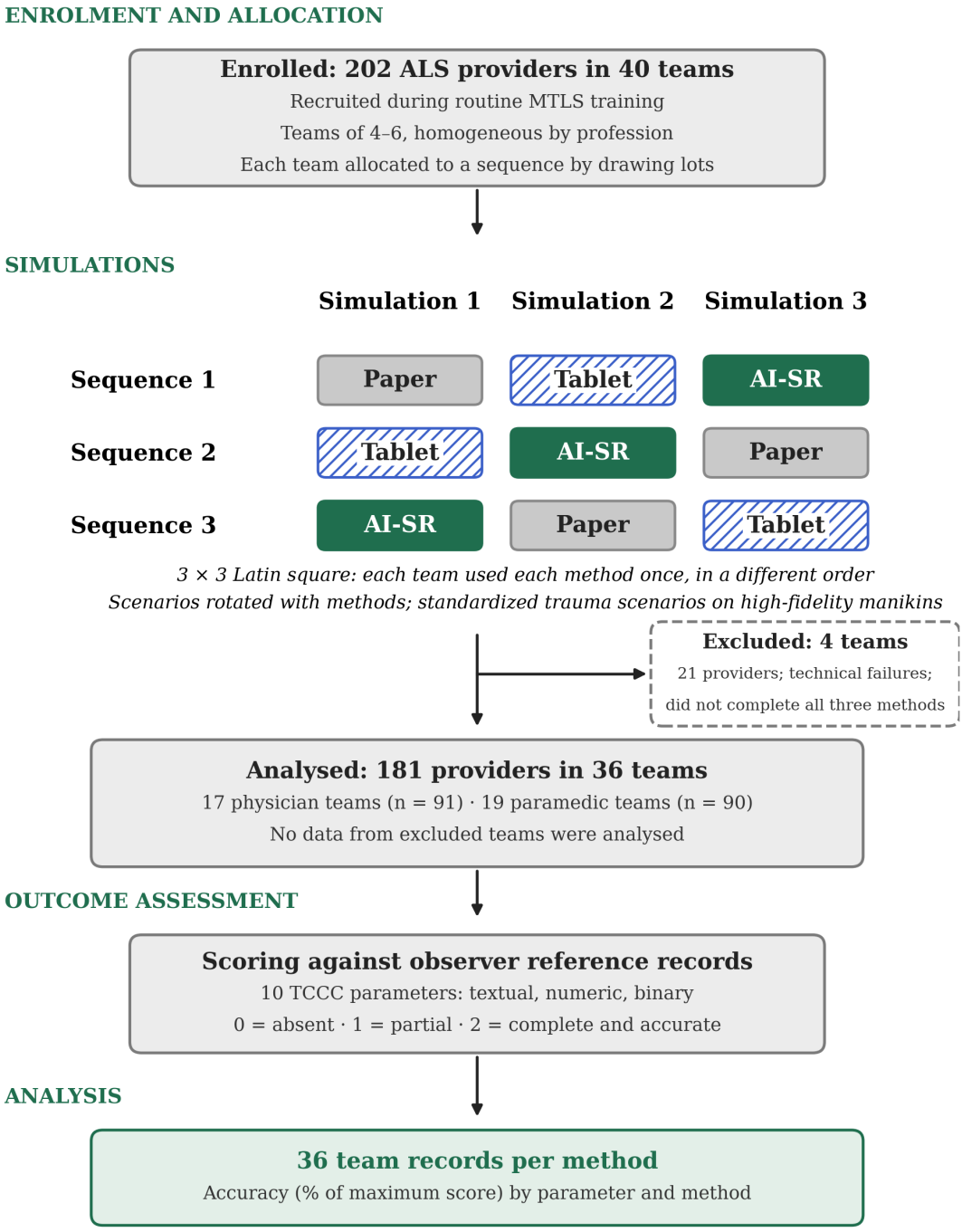


Figure 1. Study flow: enrolment, allocation, simulations and outcome assessment. Participant flow. Teams were allocated by lot to one of three sequences of a 3 × 3 Latin square; each team used each documentation method once. Four teams were excluded after allocation because of technical failures. AI-SR, artificial intelligence-based speech recognition; ALS, advanced life support; MTLS, Military Trauma Life Support; TCCC, Tactical Combat Casualty Care.

Setting and Participants

Participants were advanced life support (ALS) providers (physicians and paramedics) certified in Military Trauma Life Support, recruited during routine training sessions at the facility. Participants were organized into teams of four to six providers of the same profession (physician-only or paramedic-only); team composition remained unchanged across all three simulations. A total of 202 providers in 40 teams were enrolled and allocated to a documentation sequence. Four teams (21 providers) were excluded because of technical failures: AI-SR audio capture failed in two teams and the tablet application crashed in two teams. Because these teams did not complete documentation with all three methods, none of their data were included in the analysis. The final analytic sample comprised 181 providers (91 physicians, 90 paramedics) in 36 teams: 17 physician-only, 19 paramedic-only (Figure 1).

Simulation Protocol

Each team completed three standardized trauma scenarios on high-fidelity manikins, each lasting approximately 15 minutes: (1) head injury; (2) limb injury; and (3) torso injury. The scenarios were designed to be comparable in clinical complexity and documentation load, with a similar number of scorable elements. Scenarios were rotated together with documentation methods, so that across teams each method was used with each scenario; scenario was therefore not confounded with documentation method.

(A) Paper TCCC card: the standard form with checkboxes and free-text fields.

(B) Tablet-based system: a ruggedized Galaxy Tab touchscreen device with a structured digital interface (Samsung).

(C) AI-SR agent: a custom GPT built on GPT-4o (OpenAI, San Francisco, CA, USA), accessed through the ChatGPT web interface on participants' personal smartphones. The system was cloud-based and required an internet connection; speech was captured with the smartphones' built-in microphones, with no noise reduction beyond the devices' built-in suppression. Participants could review and edit the generated record; accuracy was scored on the final record submitted by each team, after any participant corrections.

Documentation Accuracy Measurement and Field-Type Classification

For each team, one trained instructor (36 in total) observed all three simulations without intervening and created a reference record of all patient information, injuries identified, interventions performed, vital signs obtained and medications administered, according to TCCC documentation requirements. Because the three methods produce visually distinct records, scoring could not be blinded to documentation method. Before data collection, observers participated in a joint calibration session and followed predefined scoring criteria to minimize inter-observer variability. Accuracy was assessed using a structured element-level scoring system (0 = absent, 1 = partial, 2 = complete and accurate), expressed as a percentage of the maximum possible score per field. Ten TCCC parameters were classified by structural field type:

Textual – Personal Data (PD), Injury (I), Evacuation (E): requiring free-text generation or recall of narrative clinical information.

Numeric – Safety (S), Breathing (B), Cardiovascular (C), Disability (D), Secondary Round (SR): requiring quantitative entry of physiological measurements, dosages, or structured scoring values.

Binary – Airway (A), Resuscitation Round (RR): requiring selection among discrete options or checkbox-style responses.

Statistical Analysis

The unit of analysis was the team (n = 36). The primary outcome was concordance of parameter rankings across methods; the secondary outcome was the between-method difference in accuracy within each parameter. This parameter-level analysis was pre-specified, and the sample size was set by the number of teams available, without a power calculation. Participant characteristics were compared with Mann-Whitney U and χ^2 tests. Between-parameter differences are presented descriptively. Rank concordance was assessed with Kendall's W. Within each parameter, methods were compared with Friedman tests and pairwise Wilcoxon signed-rank tests, Bonferroni-corrected for three comparisons; results were

unchanged when correcting for all 30. Effect sizes are Cohen's d_z. Tests were two-sided (α = 0.05), using R version 4.2.1.

RESULTS

Participant Characteristics

The final sample comprised 181 ALS providers: 91 physicians and 90 paramedics. Physicians were significantly older (Mdn = 31 years) than paramedics (Mdn = 20 years, P < 0.001) and more likely to be male (81.3% vs. 42.2%, P < 0.001). Prior AI-SR experience was uncommon in both groups (physicians: 14.3%; paramedics: 7.8%; P = 0.246).

Primary Finding: A Stable Parameter Hierarchy Across All Documentation Methods

Parameter rankings were highly concordant across methods (Kendall's W = 0.87, P = 0.005), and the three-tier structure was identical under all three methods (Figure 2).

A stable three-tier hierarchy emerged regardless of documentation method. The top cluster comprised S, E, and C, with overall means of 76.3%, 76.0%, and 75.6%, respectively. The middle cluster included B and A, with means of 56.4% and 55.4%, respectively. The lowest cluster included RR, PD, SR, D, and I, with overall means tightly clustered between 37.2% and 38.1%. This three-level structure (high, intermediate, and low) was consistently observed across all three documentation methods, with only minor positional shifts within clusters (Figure 2).

Specifically, within the high-performing cluster, E ranked first under paper and tablet documentation but third under AI-SR, where S achieved the highest score. Within the low-performing cluster, PD ranked sixth under paper but last under tablet, while RR was last under paper but improved relatively more under tablet and AI-SR. Despite these minor within-cluster reorderings, the three-tier structure was fully preserved across methods.

Secondary Finding: Method-Level Differences Within Each Parameter

Across all ten parameters, AI-SR consistently achieved the highest documentation accuracy, followed by tablet and paper (Figure 3). All Friedman tests were significant (all P < 0.001), and all pairwise Wilcoxon comparisons remained significant after Bonferroni correction (all P < 0.001).

The AI-SR advantage over paper ranged from 28.6 percentage points (E, d = 3.60) to 39.2 points (RR, d = 4.57). Gains were similar in the intermediate tier (A, B: Δ = 36.1–38.0 points) and the low tier (RR, PD, SR, D, I: Δ = 36.4–39.2 points), and smaller in the high tier (S, C, E: Δ = 28.6–30.4 points). The tablet system showed consistent intermediate performance, with gains of 16.4–25.3 points over paper across parameters.

DISCUSSION

This study yields two principal findings. The primary finding is that documentation accuracy in prehospital trauma care follows a stable parameter-level hierarchy preserved across documentation methods, from paper to AI-SR. The secondary finding is that AI-SR improved accuracy over both conventional methods for every parameter, with large effect sizes. Together, these findings show that, although method mattered, documentation accuracy varied substantially by parameter under every method.

This study's central result is the stability of the parameter hierarchy across methods. Parameters S, E, and C consistently achieved the highest documentation accuracy (75.6-76.3% overall), B and A clustered at intermediate levels (55.4-56.4%), and RR, PD, SR, D, and I consistently yielded the lowest accuracy (37.2-38.1%), regardless of whether providers used paper, tablet, or AI-SR. One possible explanation is that accuracy reflects unmeasured properties of each parameter: its clinical salience, the observational accessibility of the data, and the effort required to retrieve and record it. Consistent with this, readily observable parameters (safety, cardiovascular status, evacuation) were well documented under all methods, whereas those requiring measurement, calculation, or recall were poorly documented under all. Alternatively, the hierarchy may partly reflect scenario construction, which determined how many scorable elements each parameter contained and how often each was verbalized.

This finding has important implications for interpreting documentation improvement interventions. If parameter-level documentation accuracy was primarily method-dependent, then upgrading the

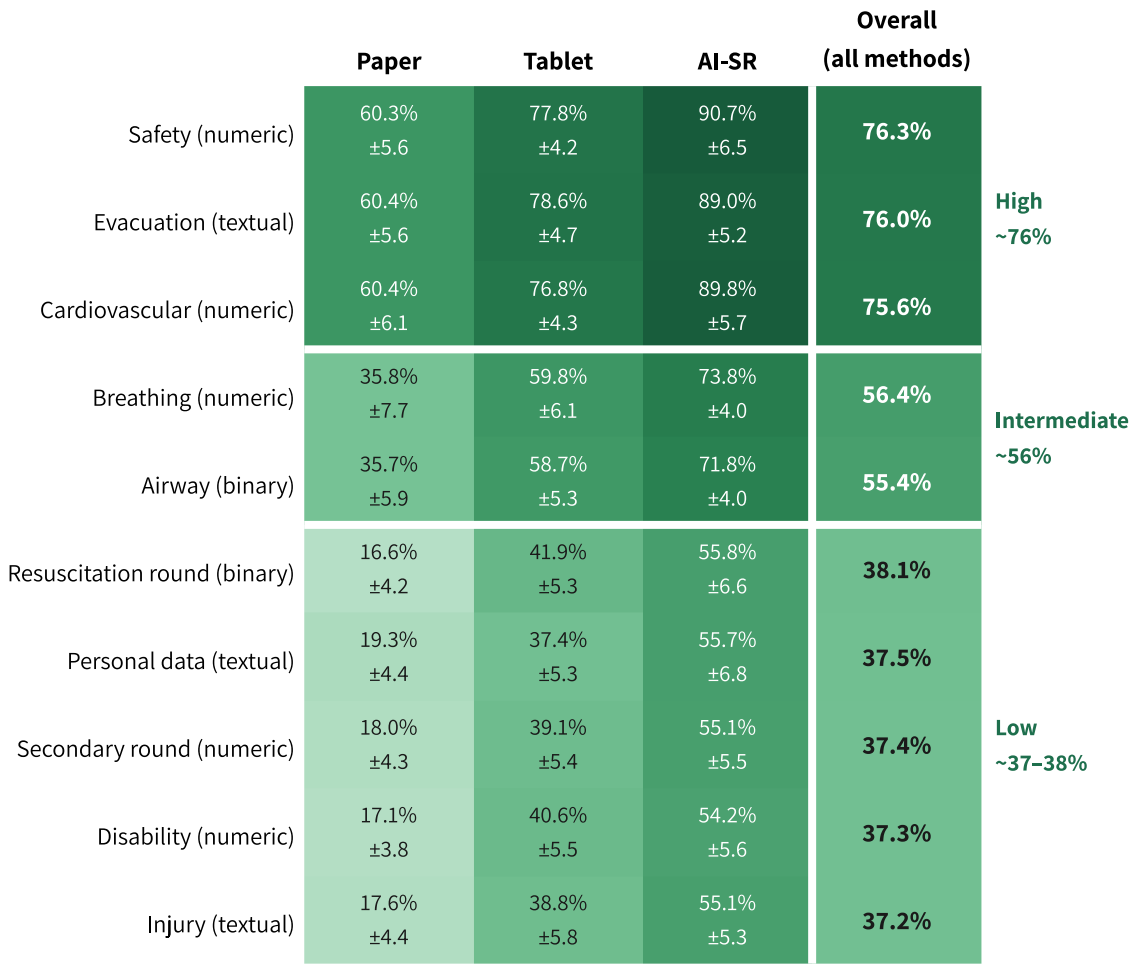


Figure 2. Documentation accuracy by parameter and documentation method in 36 simulated trauma teams. Mean accuracy (%) ± SD; n = 36 teams per method. Darker shading indicates higher accuracy; white lines separate high, intermediate, and low clusters. AI-SR, artificial intelligence-based speech recognition.

documentation interface would be expected to redistribute performance across parameters. Instead, the data suggest that upgrading from paper to AI-SR was associated with absolute gains of 28.6 to 39.2 percentage points across parameters, shifting the entire hierarchy upward. Because gains were somewhat larger in the intermediate and low tiers, the gap between the high and low tiers narrowed, from 42.6 points under paper to 34.7 points under AI-SR, but the three tiers remained distinct and rankings highly concordant (Kendall's W = 0.87). Content-level barriers to documentation may therefore persist across interfaces, and fully addressing gaps in the lowest-performing parameters may require complementary interventions, such as structured verbal prompting during care, scenario-embedded reminders, or training that targets underrepresented data elements.

The secondary finding that AI-SR consistently outperformed both conventional methods is consistent with the broader literature on hands-free documentation in time-pressured clinical environments [13–16]. AI-SR gains over paper were similar in the intermediate tier (A, B: Δ = 36.1–38.0 points) and the low tier (RR, PD, SR, D, I: Δ = 36.4–39.2 points). One possible explanation is that AI-SR captures verbalized content continuously, without requiring providers to interrupt care for manual entry; for parameters that paper captures poorly because of the cost of locating and completing the relevant field, rather than their cognitive complexity, AI-SR may remove this bottleneck. The smaller gains in the high

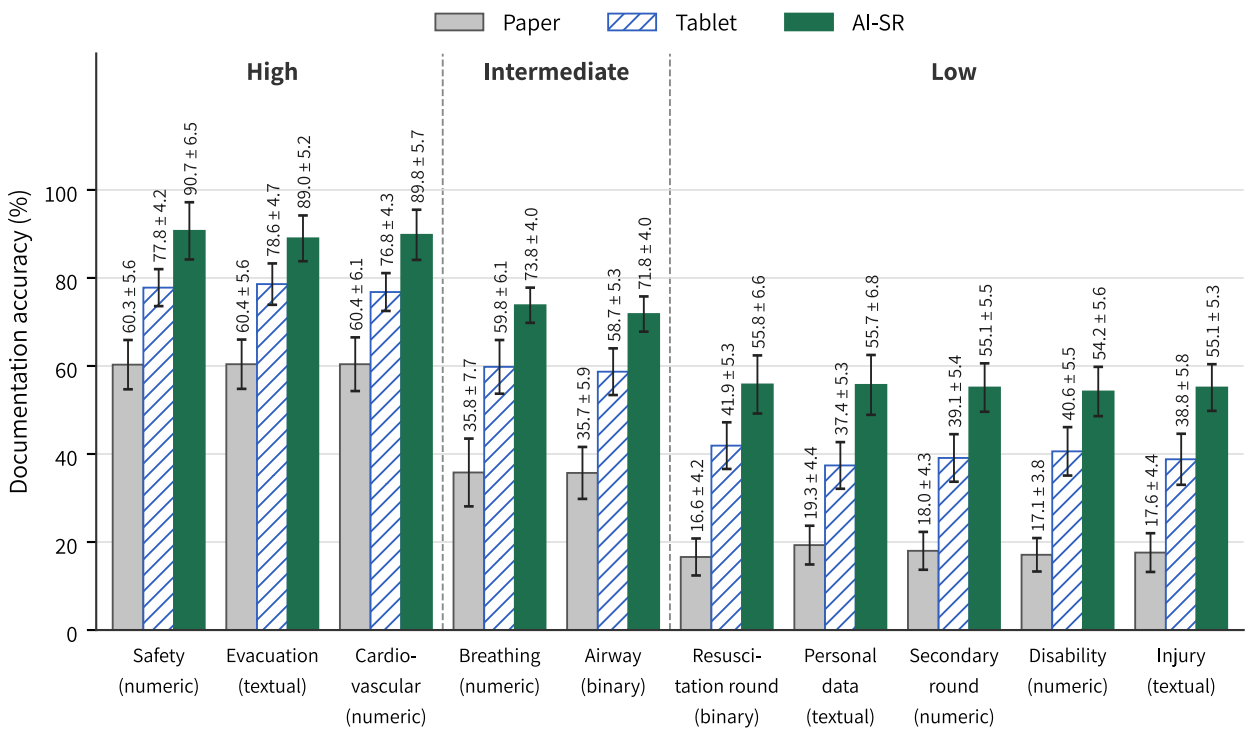


Figure 3. Mean documentation accuracy of each method within each parameter. Bars, mean accuracy; error bars and labels, mean ± SD (n = 36 teams per method). Dashed lines separate high, intermediate, and low clusters. AI-SR, artificial intelligence-based speech recognition.

tier (S, C, E: Δ = 28.6–30.4 points) may reflect a ceiling effect, as these parameters approached 90% under AI-SR. A ceiling effect does not, however, explain why the low tier, which had the most room for improvement, gained no more than the intermediate tier.

Accuracy varied widely within each field type: among textual parameters, from 76.0% (E) to 37.2% (I); among numeric parameters, from 76.3% (S) to 37.3% (D); and among binary parameters, from 55.4% (A) to 38.1% (RR). Although field type was not tested formally, this within-category heterogeneity suggests that field type alone does not explain the hierarchy, and that parameter content, rather than input format, may underlie the variation.

Limitations. Several limitations should temper the interpretation of these findings. First, and most importantly, this study was conducted under controlled simulation conditions at a single training facility, and did not replicate the acoustic, physical, and psychological stressors of an actual point-of-injury environment: background noise, gunfire, multiple simultaneous speakers, protective equipment, and genuine life-threat. Because AI-SR accuracy is directly dependent on the fidelity of speech capture, its advantage over paper and tablet documentation observed here may be attenuated in an austere or high-noise combat setting, and the present results should be regarded as an upper-bound estimate of AI-SR performance rather than a confirmed operational finding. Second, the single-facility military design limits generalizability. Third, this study did not distinguish between parameters that require immediate point-of-care documentation and those that can reasonably be completed or amended after evacuation; this distinction may itself influence both provider prioritization and susceptibility to interface-driven error, and was not modeled here. Fourth, each team completed three consecutive simulations, and we did not test whether performance changed across successive simulations independently of the documentation method used (period effects) or whether the method used in one simulation influenced performance in the next (carry-over effects). Although the Latin square design balanced the order of methods

across teams, such effects may have influenced the results. Fifth, each team's reference record was created by a single observer who was not blinded to the documentation method, and inter-rater agreement was not assessed. Sixth, four teams were excluded after allocation because of technical failures, two of them involving AI-SR audio capture; their exclusion may have led to overestimation of AI-SR performance. The stability of the parameter hierarchy across methods is therefore a descriptive finding in a simulated training environment, and its possible mechanisms require direct investigation in operational or field-training settings before these conclusions can be generalized to real-world combat casualty care.

Our scoring penalized incorrect content within required elements, but it did not penalize extraneous content absent from the reference record, such as an intervention that was discussed but not performed. AI-SR may be particularly prone to such commission errors, for example documenting a drug that was considered but not given, or mistranscribing a dose. Future evaluation of AI-SR in this setting should include a dedicated hallucination/commission-error metric alongside the accuracy measure used here. It should examine safeguards, such as mandatory provider confirmation of numeric and pharmacological entries, before considering such agents for unsupervised operational use. Relatedly, because this study used simulated scenarios and manikin patients, the AI-SR agent did not receive or process any real patient-identifying information; deploying AI-SR on real casualties would also require dedicated data-security and confidentiality safeguards for continuously transmitted audio and transcribed clinical content, which were outside the scope of this simulation-based evaluation.

CONCLUSIONS

In this simulation study, documentation accuracy in prehospital trauma care followed a three-tier parameter-level hierarchy that was highly concordant across paper, tablet, and AI-SR documentation. AI-SR improved accuracy for every parameter while largely preserving this structure, suggesting that documentation gaps in specific parameters persist regardless of the recording interface. If confirmed in operational settings, addressing gaps in the lowest-performing parameters may require content-level and training-based interventions beyond interface redesign. Accuracy varied widely within each field type. As this study was conducted in a controlled training environment, these findings should be regarded as hypothesis-generating and confirmed through field-based validation before informing operational documentation policy.

Disclaimer: The views expressed in the submitted article are the authors' own and not an official position of the IDF or the IMA.

Conflict of interest: none to declare

Source of support: none to declare

References

1. Therien SP, Nesbitt ME, Duran-Stanton AM, Gerhardt RT. Prehospital medical documentation in the Joint Theater Trauma Registry: a retrospective study. J Trauma 2011; 71(1 Suppl):S103-S108.
2. Tsur AM, Nadler R, Lipsky AM, et al. The Israel Defense Forces Trauma Registry: 22 years of point-of-injury data. J Trauma Acute Care Surg 2020; 89(2S Suppl 2): S32-S38.
3. Stiell A, Forster AJ, Stiell IG, van Walraven C. Prevalence of information gaps in the emergency department and the effect on patient outcomes. CMAJ 2003; 169(10): 1023-1028.
4. Laudermilch DJ, Schiff MA, Nathens AB, Rosengart MR. Lack of emergency medical services documentation is associated with poor patient outcomes: a validation of audit filters for prehospital trauma care. J Am Coll Surg 2010; 210(2):220-227.
5. Murad MK, Larsen S, Husum H. Prehospital trauma care reduces mortality. Ten-year results from a time-cohort and trauma audit study in Iraq. Scand J Trauma Resusc Emerg Med 2012; 20:13.
6. Lockey DJ. Research questions in pre-hospital trauma care. PLoS Med 2017; 14(7):e1002345.
7. Sandi C. Stress and cognition. Wiley Interdiscip Rev Cogn Sci 2013; 4(3):245-261.
8. Robinson JB, Smith MP, Gross KR, et al. Battlefield documentation of tactical combat casualty care in Afghanistan. US Army Med Dep J 2016; (2-16):87-94.
9. Nissley LE, Rodriguez R, April MD, et al. Occam's razor and prehospital documentation: when the simpler solution resulted in better documentation. Med J (Ft Sam Houst Tex) 2023; (Per 23-1/2/3):81-86.

10. Sawe HR, Reynolds TA, Weber EJ, et al. Trauma care and capture rate of variables of World Health Organization data set for injury at regional hospitals in Tanzania: first steps to a national trauma registry. BMC Emerg Med 2020; 20(1):29.
11. Blackley SV, Huynh J, Wang L, et al. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. J Am Med Inform Assoc 2019; 26(4):324-338.
12. Adedeji A, Joshi S, Doohan B. The sound of healthcare: improving medical transcription ASR accuracy with large language models [preprint]. arXiv 2024; arXiv:2402.07658.
13. Blackley SV, Schubert VD, Goss FR, et al. Physician use of speech recognition versus typing in clinical documentation: a controlled observational study. Int J Med Inform 2020; 141:104178.
14. Perkins SW, Muste JC, Alam T, Singh RP. Improving clinical documentation with artificial intelligence: a systematic review. Perspect Health Inf Manag 2024; 21(2):1d.
15. Hedderson D, Courtney KL, Monkman H, Blanchard IE. Speech recognition technology in prehospital documentation: a scoping review. Int J Med Inform 2025; 193:105662.
16. Woo M, Mishra P, Lin J, et al. Complete and resilient documentation for operational medical environments leveraging mobile hands-free technology in a systems approach: experimental study. JMIR mHealth UHealth 2021; 9(10):e32301.