Original Paper

Standardized Performance Assessment Methodology and Evaluation Framework for Tactical Combat Casualty Care Autonomous Documentation Algorithms: Framework Development and Evaluation

Jaeyeon Lee[1], PhD; Jeanette Little[1], BS, MS; Ethan Quist[2], ME; Omar Badawi[1], MPH, PharmD; Nathan Fisher[1], MS

[1]United States Army Institute of Surgical Research, Fort Detrick, MD, United States
[2]Defense Advanced Research Projects Agency, Arlington, VA, United States

Corresponding Author:

Jeanette Little, BS, MS
United States Army Institute of Surgical Research
Bldg. 1054 Patchel Street
Fort Detrick, MD 21702
United States
Phone: 1 7067872394
Email: jeanette.r.little.civ@health.mil

Abstract

Background: The Military Health System mandates medical documentation at all echelons of care; however, care providers in combat situations must prioritize lifesaving measures over record-keeping, leading to information gaps. Effective human-machine teaming solutions designed to autonomously document care delivery will serve as future force multipliers in tactical combat casualty care (TCCC) environments.

Objective: This study proposed a standardized performance evaluation framework for AI algorithms designed to passively document care delivery in the TCCC environment. Common AI performance evaluation methods do not adequately represent the temporal, repetitive, and context-dependent nature of real-world TCCC delivery. Comprehensive assessments are necessary to evaluate AI tools in a manner that reflects the longitudinal and workflow-dependent nature of care delivery.

Methods: To assess performance, the research team leveraged a reserved dataset. In the first phase of the assessment, a standardized, repeatable performance methodology and framework was used to evaluate individual algorithms that detect (1) treatments administered by the care provider, (2) visible medical objects, and (3) injuries and their localizations. Detection effectiveness included 4 metrics: Jaccard Index, also known as intersection over union, precision, recall, and F_1-score. Processing efficiency was evaluated by lag time scores. The second phase incorporated a centralized orchestration framework to support representative edge-constrained and synchronized execution, as well as computational resource characterization within a workflow-aware environment. In addition, quantified memory consumption and processing unit resource usage were assessed.

Results: Results are summarized for a representative algorithm in each category to illustrate the evaluation framework. Medical procedure detection yielded procedure-level mean F_1-scores from 0.00 to 0.31 and simulation-level means from 0.00 to 0.33. Higher mean F_1-scores were observed for nasopharyngeal airway, chest seal application, and treat a casualty for a cold injury (mean F_1=0.28, 0.31, and 0.31, respectively). Medical object detection achieved mean F_1-scores of approximately 0.42 (range 0-0.71). Injury detection and localization showed mean F_1-scores of approximately 0.27 (range 0-0.60). These results illustrate the ability of the framework to characterize performance variability across algorithms and simulations.

Conclusions: The proposed evaluation framework provides standardized, workflow-aware assessment of AI algorithms for core TCCC documentation tasks. While algorithm performance to date is low, the framework captures performance variability across algorithms and supports targeted identification of underperforming areas for iterative AI model refinement. The framework provides a foundation for broader evaluation of clinical and digital health systems that require event-level reasoning and workflow-aware assessment.

JMIR Med Inform 2026;14:e95075; doi: 10.2196/95075

Keywords: tactical combat casualty care; autonomous documentation algorithms; performance assessment methodology; injury detection; injury location; medical object detection; medical procedure detection

Introduction

The Military Health System faces a persistent challenge: while medical documentation is mandated at all echelons [1], providers in high-intensity combat situations must prioritize lifesaving measures over record-keeping, leading to significant data gaps. The US military is prioritizing human-machine teaming (HMT) solutions focused on integrating AI and autonomous systems to improve efficiency, speed, and decision-making in the operational environment [2]. With respect to tactical combat casualty care (TCCC), the development of passive, automated HMT resources for care documentation addresses a gap in austere environments. Recent studies have further emphasized the growing role of AI-enabled HMT and decision support systems in military medicine and operational health care environments [3,4].

To address this challenge, the United States Army Institute of Surgical Research (USAISR), formerly known as the Telemedicine and Advanced Technology Research Center, commenced an effort to automate the care documentation process through passive modalities. The premise behind this effort was to prototype solutions capable of autonomous computation and data exchange [5] that offer a unique solution to this TCCC medical documentation capability gap. By integrating technologies, specifically their underlying AI algorithms, the resulting HMT tools will serve as a force multiplier for the future TCCC-level provider, including the following:

1. Reduced cognitive load: By automating the capture of patient statuses and caregiver actions, the system alleviates the medic's cognitive burden, allowing them to remain fully "hands-on" with the patient without the distraction of manual documentation [6,7].
2. Data integrity for handoffs: The system facilitates the accurate generation of digital TCCC care documentation, ensuring that life-critical information (including but not limited to medication dosages and tourniquet application times) is seamlessly transferred to the next echelon of care without requiring the medic to pause treatment for manual logging [8].
3. Mass casualty and patient management: Mature technologies will enable continuous, simultaneous, and passive monitoring of multiple vital signs across multiple patients through a centralized interface, significantly enhancing a care provider's situational awareness in complex environments [9]. Recent reviews have also highlighted the increasing use of AI-assisted monitoring and wearable sensing technologies across trauma and emergency care workflows [10].
4. Enhanced operational readiness: Near real-time monitoring of physiological parameters and environmental exposures (eg, noise, radiation, and chemical agents) ensures that the health and safety of military members are maintained throughout the mission [11-13].

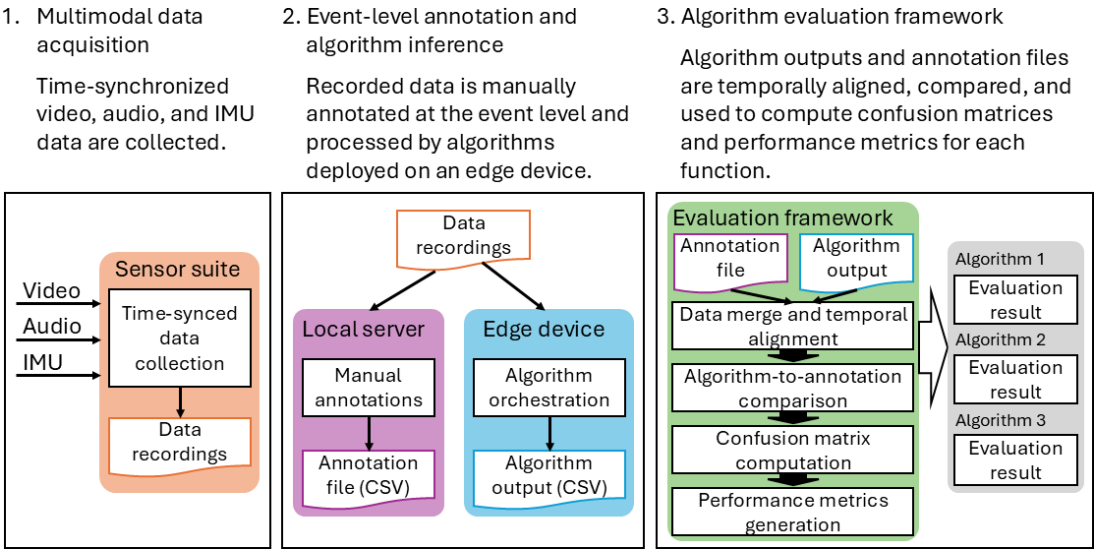
In May 2023, the USAISR commenced the effort to prototype HMT systems designed to passively document care delivery within the TCCC environment. This overall autonomous documentation project has four main hypotheses: (1) Synchronized passive sensor modalities can establish large, reliable TCCC encounter datasets. (2) Passively collected data can effectively train future patient documentation algorithms. (3) Mature algorithms can reliably automate real-time data inputs for digital TCCC Cards. (4) Passive, autonomous collection will eliminate the need for manual care provider input at the point of care, improving both efficiency and accuracy.

The overall approach to research-based prototype development and evaluation was multifaceted, as shown in Figure 1. Initially, passive sensor components, including video, audio, and activity tracking [14], captured time-synchronized data associated with both TCCC provider actions and engagement with injured Service Members. The data generated by these inputs are processed via 2 complementary pathways. First, the data are curated by a government laboratory and manually annotated at the event level by TCCC experts. Second, the data are processed on an edge-computing device, where multiple AI algorithms are orchestrated to detect treatments administered by the care provider, medical objects visible in the scene, and injury patterns and locations present on the casualty.

Synchronized outputs from these algorithms are then consolidated to produce passive and autonomous TCCC documentation. During the evaluation stage, the algorithm outputs and event-level annotations are merged and temporally aligned, followed by algorithm-to-annotation comparison, enabling confusion matrix computation and performance metric generation for algorithm evaluation.

To initiate the project, a robust TCCC encounter dataset was established using medical simulation under a human subject research protocol; it has been described in previous peer-reviewed publications [15,16]. Capturing accurate and robust TCCC data was a critical step in this project, as it enabled the development of the critical AI resources to automate the HMT solution [17].

Figure 1. Research project approach. Overview of the research workflow from multimodal data acquisition and event-level annotation to algorithm inference and performance evaluation. CSV: comma-separated value; IMU: inertial measurement unit.



Once established, the TCCC dataset was annotated by the government research team members with event-based metatags that characterized TCCC provider interventions. These descriptors were then used as ground-truth labels for algorithm model training and evaluation. An example of the government-generated annotations is shown in Figure 2.

The annotations provide a high-level overview of casualty treatments. Each entry specifies the medical procedure (Task) performed by the caregiver, its start and end times (Skill Start Time and Skill End Time), and the total duration of the procedure (Time on Task). The anatomical context of each intervention is captured through the body, anatomical, and sagittal location fields. Additional annotation categories document procedural details, including administered medications, dosage, and administration route when observable. The abandonment column indicates when medical procedures were discontinued prior to completion, and the notes provide contextual information, such as simulation conditions or other relevant observations. These event-level annotations, generated through expert review by trained TCCC experts, defined the clinical reference used for algorithm performance evaluation. The provided annotations were limited to event-level labels and did not include fine-grained annotations such as bounding boxes or frame-level labels.

Figure 2. Tactical combat casualty care dataset annotation example. Example of event-level annotations used for algorithm training and evaluation, including procedure timing, anatomical location, medication information, and procedure abandonment.

Task	Skill Start Time*	Skill End Time*	Time on Task*	Body Location	Anatomical Location	Sagittal Location	Medication	Dosage	Route	Abandonment (Y)	Notes
STARTEX	0:00:00	0:00:35	0:00:35								
Time Sync	0:00:35	0:00:35	0:00:00								
Junctional Wound Packing	0:01:26	0:02:45	0:01:19	Thigh	Anterior	Left					
Pressure Dressing Application	0:02:45	0:31:07	0:28:22	Thigh	Anterior	Left				Y	No stop cue reached.
Surgical Cricothyrotomy	0:04:29	0:07:11	0:02:42	Neck							
BVM Assisted Ventilation	0:06:28	0:06:38	0:00:10	Neck							
Uses Sensor	0:07:24	0:07:30	0:00:06	Hand							
Treat a Casualty with a Pelvic Fracture	0:09:00	0:10:04	0:01:04	Pelvis							
Initiate an IV	0:10:45	0:13:26	0:02:41	Lower Arm							
Manage Intravenous Access	0:12:57	0:13:15	0:00:18						Intravenous		
Administer Medication	0:13:48	0:17:01	0:03:13	Lower Arm			TXA	2g	Intravenous		
Treat a Casualty for a Cold Injury	0:17:18	0:31:07	0:13:49							Y	No stop cue reached.
Documentation	0:18:04	0:18:44	0:00:40								
Uses Sensor	0:18:51	0:18:59	0:00:08								
Documentation	0:19:03	0:19:43	0:00:40								
Documentation	0:21:02	0:21:13	0:00:11								
Documentation	0:21:18	0:21:26	0:00:08								
Administer Medication	0:21:30	0:22:21	0:00:51	Lower Arm			Ketamine	10mg	Intravenous		
Administer Medication	0:22:25	0:23:33	0:01:08	Lower Arm			Ertapenem	1g	Intravenous		
Documentation	0:23:40	0:24:17	0:00:37								
Documentation	0:26:53	0:30:40	0:03:47								
Uses Sensor	0:24:26	0:24:30	0:00:04								
Documentation	0:30:48	0:30:52	0:00:04								
ENDEX	0:30:53	0:31:07	0:00:14								

*Time is presented in a 24 hour clock format of HH:MM:SS where HH=hours, MM=minutes and SS=seconds

Once the annotation process had been completed, a subset of this dataset was provided to algorithm developers for model training and development, while another portion was held in reserve for independent performance assessment. In January 2025, 5 algorithm development partners were provided a subset of the TCCC dataset consisting of 75 unique TCCC simulations based on standardized scenarios to develop their documentation algorithms using the various techniques they proposed.

To assess these heterogeneous algorithms in a consistent and unbiased manner, USAISR needed a method for independently assessing the performance of delivered

documentation algorithms using a reserved evaluation dataset. This method had to be independent of the various models and algorithm training methods used by each individual development team and repeatable during subsequent algorithm development cycles as more data from the TCCC dataset were published.

While multialgorithm documentation systems allow continuous capture of care delivery, they also make performance evaluation more challenging. Although autonomous medical vision techniques have advanced significantly, there remains limited focus on developing standardized approaches for assessing algorithm performance in complex, longitudinal clinical settings in TCCC. Common evaluation methods, often adapted from static classification tasks, rely on summary metrics (eg, sensitivity, specificity, or positive predictive value) that do not adequately represent the temporal, repetitive, and context-dependent nature of real-world care delivery [18,19].

In TCCC scenarios, algorithms produce continuous outputs over extended periods. Multiple detections may represent ongoing clinical states rather than separate events, and delayed identification of time-critical interventions can reduce clinical and operational usefulness. Conventional evaluation approaches do not always account for the temporal characteristics of continuous algorithm outputs or the event-level objective of longitudinal clinical documentation, particularly when the operational goal extends beyond frame-level event localization. Evaluation strategies designed for temporally constrained or frame-level assessment do not necessarily align with the objective of autonomous TCCC documentation, where the primary goal is to document clinically meaningful events throughout the overall treatment process rather than to localize every detection at the frame level [20]. Therefore, event-level evaluation provides a more appropriate framework for autonomous TCCC documentation by reflecting the intended clinical documentation workflow while remaining compatible with the level of annotation currently available [21]. The aim of this study was to develop a standardized performance assessment methodology for autonomous documentation algorithms operating within a workflow-aware TCCC operational framework.

Methods

Study Design

To enable reliable deployment of multiple algorithms on a resource-constrained edge device, algorithm inputs were constrained by fixed time windows, batch sizes, and tensor shapes. These constraints ensured stable execution, predictable resource usage, and compatibility with the orchestration software framework, while also supporting consistent performance assessment. Table 1 summarizes the input data requirements. Video was sampled in 1-second cycles (30 frames at 30 frames per second), audio in 30-second segments of 16 kilohertz mono pulse-code modulation totaling 480,000 samples, and inertial measurement unit data in 1-second windows as 6 degrees of freedom signals (eg, acceleration and gyroscope on x-, y-, and z-axes) collected at 60 hertz from sensors on both wrists (eg, left and right).

Table 1. Algorithm input requirements for various modalities.

Modality	Time (seconds)	Batch size, n	Tensor	Description
Video	1	30	(30, 3, 416, 416)	Represents 1 second of video
Audio	30	1	(480,000)	16-kHz mono audio, PCM[a] 32-bit floating-point big-endian
IMU[b]	1	60	(60, 12)	60 Hz (ACC[c] x, y, z, L/R[d]; Gyr[e] x, y, z, L/R)

[a]PCM: pulse-code modulation.
[b]IMU: inertial measurement unit.
[c]ACC: acceleration.
[d]L/R: left and right.
[e]Gyr: gyroscope.

Algorithm outputs were required to align with the fields defined in the TCCC card, Department of Defense Form 1380, while supporting objective and standardized performance evaluation. Output requirements for different algorithm functionalities are summarized in Table 2. To ensure time synchronization and interoperability across modules, each output was required to include the pipeline name, the absolute time stamp in International Organization for Standardization 8601 format, the elapsed time and corresponding frame index obtained from the source video, and a patient identifier for cases involving multiple patients.

For procedure and object detection, algorithm outputs were used to populate the injury and treatments fields of the TCCC Card, including the predicted procedure and the detected medical object with their respective confidence scores. The medical procedures and associated objects that were identified and prioritized for algorithm development within the proposed evaluation framework are highlighted in Table 3.

Table 3 lists the complete set of 25 targeted medical procedures. However, at the current stage, only 13 procedures are covered by the available training dataset. Therefore, the representative algorithms were trained and evaluated on these 13 procedures, and the results presented in this paper are limited to this subset. The remaining medical procedures will be included in future evaluations as additional training data becomes available and the algorithms are further developed.

Table 2. Output requirements for different algorithm functionalities.

Algorithm functions	Output requirements
Medical procedure detection	• Procedure • Procedure confidence score
Medical object detection	• Object • Object confidence score
Injury detection and localization	• Injury detection and localization • Injury detection confidence score • Injury location confidence score • Injury pattern[a] • Injury pattern confidence score[a]

[a]The injury pattern and its associated confidence score will be implemented in future phases of the tactical combat casualty care algorithm assessments.

Table 3. TCCC[a] medical procedures and objects.

Medical procedure	Medical objects
Administer fluids through an infusion	500-mL NS[b] IV[c] bag; IV tubing set
Administer medication	3-mL syringe; 10-mL saline flush; 25-gauge needle; blunt-tip needle; training medication (simulated tablet, capsule, liquid, or topical medication; simulated nasal drops/spray; and simulated inhaler or nebulizer)
Administer whole blood	Training packed red blood cell unit; blood administration set
BVM[d]-assisted ventilation	Cyclone BVM
Chest needle decompression	North American Rescue Air Release System, 14-gauge
Chest seal application	Chest seal
Extraglottic airway	iGel superglottic airway device (large adult)
Initiate an IV access	Chinook Medical Gear IV Start Kit
Junctional tourniquet application	SAM[e] Junctional Tourniquet
Junctional wound packing	Combat Gauze
Manage a minor laceration	Emergency trauma bandage
Manage IV access	10-mL saline flush
Nasopharyngeal airway (NPA)	14-French NPA; lubricant/simulated lubricant
Operate a fluid warmer	Fluid warmer
Oropharyngeal airway	Berman oral airway
Perform patient suctioning	Manual suction device
Place an IO[f] device	10-mL NS flush; EZ-IO needle
Pressure dressing application	Emergency trauma bandage
SAVE[g] II ventilator operation	SAVE II ventilator
Splint a casualty with extremity injury	Structural aluminum malleable splint; ACE[h] wrap
Surgical cricothyroidotomy	Tactical cricothyroidotomy kit (CricKit)
Tourniquet application	Combat Application Tourniquet
Treat a casualty for a cold injury	Hypothermia Prevention and Management Kit
Treat a casualty with a pelvic fracture	SAM pelvic binder
Treat a casualty with burns	Kerlix antimicrobial dressing

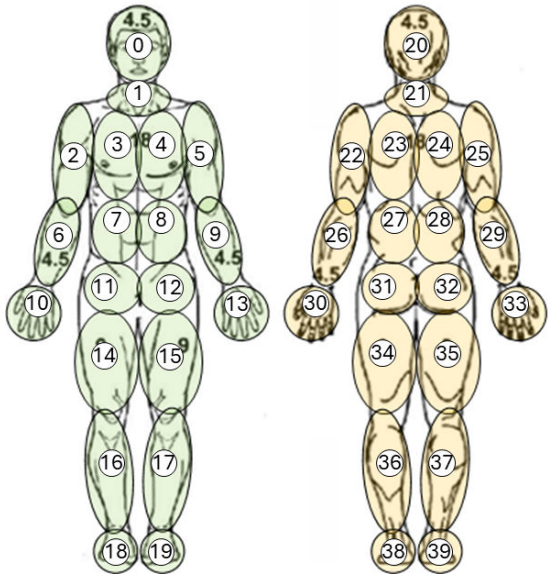
[a]TCCC: tactical combat casualty care.
[b]NS: normal saline.
[c]IV: intravenous.
[d]BVM: bag valve mask.
[e]SAM: Structural Aluminum Malleable.
[f]IO: intraosseous.
[g]SAVE: Simplified Automated Ventilator.
[h]ACE: All Cotton Elastic (as in ACE bandage).

Outputs from injury detection and localization algorithms populated the injury fields of the TCCC Card, specifying whether an injury was detected, its location according to a standardized anatomical schema, the relevant confidence score, the injury pattern, and its associated confidence score. Reported injury locations were constrained to predefined anatomical regions, as illustrated in Figure 3.

Algorithm evaluation was performed using a standardized evaluation framework (Python, version 3.10; Python Software Foundation) on a Linux (Ubuntu, version 20.04)

environment. The evaluation process began by preprocessing TCCC annotation and algorithm output files into a standardized comma-separated value format. Time stamps were used to synchronize ground-truth labels and algorithm outputs. Algorithm outputs were then matched to ground-truth annotations using task-specific event-matching criteria based on annotated procedure intervals, simulation-level event presence, and predefined anatomical mappings. Finally, confusion matrices were computed and aggregated to generate performance metrics. Results were exported as text files, and intermediate outputs were retained for reference and traceability. Adjustable parameters provided via a configuration file enabled flexible customization to accommodate algorithm-specific requirements.

Figure 3. Casualty injury location body mapping. Standardized body region mapping used for injury detection and localization evaluation. ID: identifier.



ID, body region	
0, Front_Head	20, Back_Head
1, Front_Neck	21, Back_Neck
2, Front_Right_Upper_Arm	22, Back_Left_Upper_Arm
3, Front_Right_Upper_Torso	23, Back_Left_Upper_Torso
4, Front_Left_Upper_Torso	24, Back_Right_Upper_Torso
5, Front_Left_Upper_Arm	25, Back_Right_Upper_Arm
6, Front_Right_Lower_Arm	26, Back_Left_Lower_Arm
7, Front_Right_Lower_Torso	27, Back_Left_Lower_Torso
8, Front_Left_Lower_Torso	28, Back_Right_Lower_Torso
9, Front_Left_Lower_Arm	29, Back_Right_Lower_Arm
10, Front_Right_Hand	30, Back_Left_Hand
11, Front_Right_Pelvis	31, Back_Left_Pelvis
12, Front_Left_Pelvis	32, Back_Right_Pelvis
13, Front_Left_Hand	33, Back_Right_Hand
14, Front_Right_Upper_Leg	34, Back_Left_Upper_Leg
15, Front_Left_Upper_Leg	35, Back_Right_Upper_Leg
16, Front_Right_Lower_Leg	36, Back_Left_Lower_Leg
17, Front_Left_Lower_Leg	37, Back_Right_Lower_Leg
18, Front_Right_Foot	38, Back_Left_Foot
19, Front_Left_Foot	39, Back_Right_Foot

The evaluation dataset consisted of 15 reserved TCCC simulations that were intentionally separated from the 75 simulations distributed for algorithm development to support independent performance evaluation. The evaluation dataset included 6 simulations conducted using manikins and 9 simulations involving live simulated patients. It included diverse operational conditions across indoor and outdoor TCCC-relevant environments, including laboratory and tent settings, standardized lighting conditions, a low-light, red illumination environment, and outdoor daylight conditions. All simulations were performed by trained expert medics. Additional details regarding the TCCC data collection protocols, simulated operational environments, sensor configurations, and live simulated patient data collection events are described in prior publications [14,15]. The evaluations conducted to date have focused on single-patient scenarios using video-based data collected during TCCC simulations.

Ground-truth annotations were generated from TCCC simulation recordings using event-level labels aligned with TCCC documentation requirements. Medical procedure annotations included start and end times, while medical object presence was derived from the procedures performed during the simulation rather than from direct temporal object annotations. Injury localization annotations identified the anatomical location of injuries observed during the simulation. Additional details regarding the annotation structure and labeling methodology are available in separate publications [15]. For each simulation recording, annotations were independently generated by multiple expert annotators and subsequently reviewed and consolidated into a finalized annotation set. Ambiguous labels, incomplete intervals, and inconsistent time stamp entries were reviewed during this process to ensure compatibility with the standardized evaluation framework.

As part of the annotation quality control, agreement between independent annotations was evaluated based on the percentage of temporal overlap for each annotated task. Temporal overlap was calculated using intersection over union, defined as the number of overlapping annotated time points divided by the total number of unique annotated time points from both annotators. Across the available annotation records from 82 simulations, a total of 586 task comparisons were available after excluding the nontask StartEX, EndEX, and Time Sync entries. These comparisons involved 3 distinct annotator pairs formed by a total of 4 annotators. The 82 simulations included 6 evaluation simulations conducted using manikins. The remaining 9 evaluation simulations involved live simulated patients and were collected, annotated, and quality-checked under a separate protocol. They were not included in the overlap analysis. The analysis did not include tasks recorded by only 1 annotator. This means that the overlap results reflect agreement in annotation timing for tasks recorded by both annotators, rather than a separate measure of task identification agreement.

Before adjudication, the median overlap was 96.4%, and 448 of 586 (76.5%) comparisons met the predefined 80% overlap rule [15]. The remaining 138 of 586 (23.5%) comparisons that did not meet the overlap rule were further

reviewed and adjudicated. During this process, disagreements in annotation timing, task labels, abandonment status, and task inclusion were reviewed and resolved. After adjudication, all 586 comparisons (100%) met the 80% overlap requirement, with a median overlap of 98.4%. These results show the outcome of the annotation quality control process rather than independent agreement between annotators.

The government laboratory performed a 2-part assessment to evaluate these algorithm deliverables. In the first phase, each individual algorithm was assessed independently using the proposed performance metrics in an algorithm-centric manner. In the second phase of the evaluation, algorithms were assessed in an orchestrated, system-level manner under realistic edge-computing constraints. Algorithm performance was evaluated through individual execution under standardized resource constraints on a single edge-computing device, while system-level resource usage was evaluated through concurrent execution of multiple algorithms on the same device.

Phase 1: Individual Assessments of TCCC Patient Documentation Algorithms

In the first phase of the TCCC patient documentation algorithm assessment, each individual algorithm was evaluated against standardized performance metrics on a high-performance development computing device. Algorithms were provided as stand-alone software packages, capable of independent execution in a standard development environment, and were assessed as single software instances. Three categories of algorithms were considered: (1) medical procedure detection, (2) medical object detection, and (3) injury detection and localization. This evaluation focused on two primary dimensions: (1) detection effectiveness, with Jaccard Index, precision, recall, and F_1-score; and (2) processing efficiency, assessed using the lag time (LT) score.

The evaluation used confusion matrices based on true positives (TPs), false positives (FPs), and false negatives (FNs) to account for the limited availability of ground-truth annotations in the TCCC dataset, the continuous and longitudinal nature of algorithm outputs, and operational characteristics of the algorithms. Because true negatives are difficult to define in this setting, where algorithm outputs are continuous and context-dependent, the conventional accuracy formula was not applicable. Instead, the Jaccard Index was adopted, and precision, recall, and F_1-score were computed from TP, FP, and FN counts specifically defined based on the operational characteristics as shown in Table 4.

In this framework, recall is calculated at the algorithm-output level. Since FNs represent empty or invalid outputs rather than all annotated items missed by the algorithm, recall does not measure annotated item coverage. To address this, ground-truth coverage was calculated as a complementary measure of detection effectiveness for each detection category based on the available ground-truth annotations. For procedure detection, coverage was calculated as the percentage of annotated procedure events when 1 or more TP detections occurred during the corresponding annotated time intervals. For medical object detection, coverage was calculated as the percentage of procedure-derived medical objects when 1 or more TP detections occurred during the simulation. For injury detection and localization, coverage was calculated as the percentage of annotated anatomical regions when 1 or more TP detections occurred during the simulation. Each annotated procedure event was counted separately, while each medical object or anatomical region was counted only once per simulation, regardless of the number of TP detections.

Table 4. Tailored definitions by algorithm function[a].

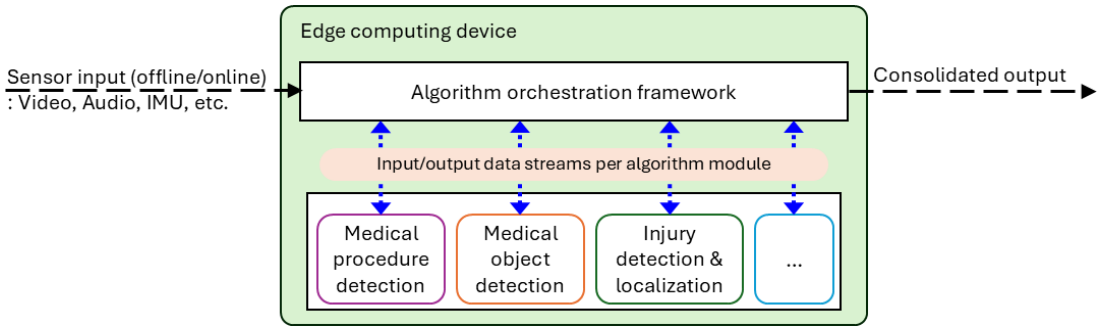
Algorithm function	TP[b]	FP[c]	FN[d]
Medical procedure detection	An algorithm output that matches the annotated procedure occurring within the corresponding annotated time interval.	An algorithm output that does not match the annotated procedure within the corresponding time interval, ie, an output generated when the predicted procedure label does not match the annotation for that interval.	An algorithm output that fails to produce any valid procedure detection for a given procedure interval, including cases recorded as [], "no detection," not a number (NaN), or other forms of detection failure.
Medical object detection	An algorithm output that matches an annotated device appearing in the set of active devices across the entire length of the simulation.	An algorithm output that does not match any annotated device, ie, an output generated when the predicted device label does not correspond to any active devices across the entire simulation.	An algorithm output that fails to produce any valid device detection, including cases recorded as [], "no detection," NaN, or other forms of detection failure.
Injury detection and localization	An algorithm output that matches the annotated injury location corresponding to any annotated body region across the entire length of the simulation.	An algorithm output that does not match any annotated injury location within the mapped body regions, ie, an output generated when the detection is reported at a location not included in the annotations.	An algorithm output that fails to produce any valid injury location or that indicates a location outside the defined body regions, including cases recorded as [], "no detection," NaN, "out of body," or other forms of detection failure.

[a]Recall is calculated at the algorithm-output level using the TP and FN definitions. It does not measure the proportion of all annotated procedures, medical devices, or injury regions detected by the algorithms.
[b]TP: true positive.
[c]FP: false positive.
[d]FN: false negative.

Phase 2: Assessment of TCCC Patient Documentation Algorithms for Timely Synchronization and Orchestration

The second phase of the evaluation focused on system-level assessment under orchestration framework constraints representative of centralized deployment environments. This phase evaluated algorithm performance under shared hardware resource constraints representative of edge deployment environments. In contrast to the algorithm-centric assessments conducted in phase 1, this phase emphasized the feasibility of coordinating multiple algorithms within a resource-constrained operational environment. The overall orchestration process is illustrated in Figure 4. For algorithm performance evaluation, only 1 algorithm was executed at a time under standardized resource constraints designed to reflect deployment on a single edge-computing device. These constraints included fixed limits on frame rate, input resolution, batch size, and shared memory usage, enabling assessment of each algorithm's ability to operate reliably under resource constraints designed for a multialgorithm edge environment. For the system-level resource usage benchmark, multiple integrated algorithms were executed concurrently on the same edge-computing device. This concurrent execution was used to evaluate the combined resource demands of multiple algorithms and their pipelines under the intended deployment configuration.

Figure 4. Phase 2 algorithm orchestration process. Overview of the orchestration framework for managing multiple algorithm modules and their input and output data streams on a single edge-computing device. IMU: inertial measurement unit.



An algorithm orchestration framework was developed to manage execution, timing control, and resource management for multiple individual algorithms operating under edge-computing constraints. The framework enables parallel online execution while optimizing computational efficiency based on the capabilities of the edge-computing device (eg, NVIDIA Jetson Orin AGX; NVIDIA Corporation). Communication between the orchestration layer and individual algorithms was facilitated via ZeroMQ (ZMQ), whereas shared memory and the GStreamer library were used to distribute intermediate outputs and support real-time processing of complex multimedia workflows. The framework acquires multimodal sensory data, applies algorithm-specific preprocessing, and routes the data to the corresponding algorithms. While individual algorithms can generate outputs independently through the orchestration framework, the orchestration layer supports aggregation and synchronization of algorithm outputs suitable for TCCC patient documentation and integration with external systems.

In this phase, the evaluation primarily focused on the performance of each algorithm module operating within the orchestration framework, rather than the consolidated outputs. This enabled precise characterization of computational efficiency, timing, and resource usage of each algorithm under edge-computing conditions, which are more constrained than the high-performance environment used in phase 1. By evaluating outputs prior to aggregation, the assessment supports algorithm module-level performance analysis.

Algorithm Evaluation Strategy

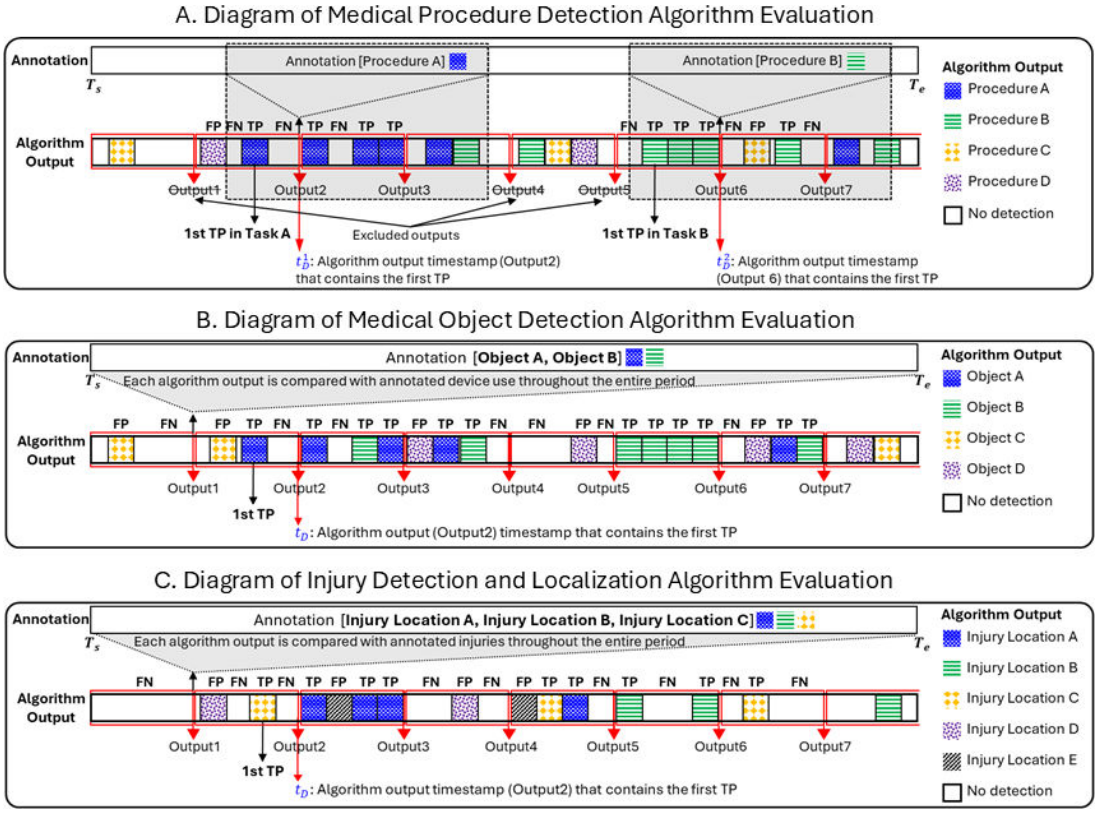
Algorithm performance was evaluated using a study-specific standardized methodology designed to ensure reproducibility and to enable comparison across diverse approaches. Confidence scores produced by the algorithms were incorporated independently of the evaluation framework and supported configurable thresholding. However, confidence-based threshold filtering was not applied in this phase due to the early stage of algorithm development. Instead, confidence scores were used only during postprocessing to consolidate duplicate outputs and select representative detections within grouped output sequences. Configurable thresholding capabilities were retained within the framework for future evaluations as algorithm maturity improves. Evaluation was conducted along 2 complementary dimensions: detection effectiveness and processing efficiency.

Figure 5 provides a diagrammatic representation of the evaluation methodology for the 3 categories of algorithm functions. Medical procedure detection is illustrated in Figure 5A, medical object detection in Figure 5B, and injury detection and localization in Figure 5C. In these diagrams, confusion matrices are computed to evaluate detection effectiveness, and time stamps for calculating LT scores are provided to assess processing efficiency.

As illustrated in Figure 5, algorithm outputs were evaluated using task-specific event-matching strategies designed for longitudinal TCCC documentation workflows. Because participating development teams implemented algorithms with different inference frequencies and output

behaviors, direct comparison of all raw outputs could bias confusion matrix statistics. Therefore, task-specific postprocessing and output grouping procedures were applied before confusion matrix computation to support balanced comparisons across heterogeneous algorithms.

Figure 5. Diagram of algorithm evaluation workflows. Evaluation workflows for (A) medical procedure detection, (B) medical object detection, and (C) injury detection and localization, illustrating the comparison of algorithm outputs with the corresponding annotations for performance evaluation. FN: false negative; FP: false positive; TP: true positive.



For all categories, outputs were compared against corresponding ground-truth annotations according to the event-level matching criteria. Outputs generated outside clinically relevant annotated intervals were excluded from evaluation. For medical procedure detection, outputs were evaluated within annotated procedure-specific start and end intervals. Consecutive repeated detections of the same procedure call were grouped, and the confidence scores were used to select representative outputs within each grouped detection sequence. Specifically, up to 5 repeated outputs with the highest confidence scores within a continuous detection sequence were retained for evaluation. This approach reduced the influence of highly repetitive outputs while preserving meaningful performance differences across procedure detection algorithms.

For medical object detection and injury detection and localization, evaluation was conducted at the simulation level rather than within the procedure-specific temporal intervals. This approach reflects the event-level documentation objective of the framework, where identifying clinically relevant medical objects or injury locations during the simulation was prioritized over exact detection timing.

To address the substantially higher output frequencies produced by the medical object detection algorithms, outputs were grouped into 30-second intervals, and duplicate detections of the same object class within each interval were consolidated, with confidence scores used to select the representative output within each grouped interval by retaining the highest-confidence detection. The same postprocessing approach was applied to injury detection and localization, where consecutive repeated outputs of the same class were consolidated by retaining the highest-confidence detection within each grouped detection sequence.

Collectively, these procedures were designed to reduce bias caused by heterogeneous output rates while maintaining clinically meaningful, event-level assessment of passive TCCC documentation algorithms. Based on these event-matching and postprocessing procedures, confusion matrices were computed to quantify detection effectiveness.

Four evaluation metrics, Jaccard Index, precision, recall, and F_1-score, were derived according to equations 1-4:

$$Jaccard\ Index = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (3)$$

$$F_1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

In addition, ground-truth coverage was assessed separately as a complementary measure of detection effectiveness to quantify the proportion of annotated items with at least 1 TP detection. Processing efficiency was evaluated as a complementary measure of algorithm performance using the time stamps of TP detections to calculate LT scores. The general approach for computing these scores is summarized in equation 5. For medical procedure detection, T_n was the total duration of the n-th procedure (procedure n), and t^n_D was the time of the first TP event on procedure n. For the remaining categories, T_n represented the total simulation duration, and t^n_D was calculated for the entire simulation to compute the LT score.

$$LT = \frac{T_n - t_D^n}{T_n}, \quad n \in \begin{cases} \{\text{Procedures}\}, & \text{medical procedure detection} \\ \{\text{Simulation}\}, & \text{others} \end{cases} \quad (5)$$

Medical Procedure Detection

Overview

For evaluation purposes, overlapping procedure annotations were handled under the assumption that only 1 procedure was active at a given time. When annotated periods overlapped, the preceding procedure was truncated at the onset of the subsequent procedure. This approach was adopted to define a single active procedure interval for standardized event-level evaluation under the current framework. Since the proposed framework evaluates whether each procedure was correctly identified at the event level rather than focusing on the precise timing of individual detections, this approach provided consistent evaluation intervals for objective and reproducible performance assessment. However, truncating overlapping annotations may affect procedure-level evaluation and represents a limitation of the current approach.

Evaluation results were generated at both the simulation and individual procedure levels, enabling performance assessment for each simulation as well as for each procedure. Algorithm outputs were evaluated within each annotated procedure interval to assess detection effectiveness for individual procedures, and metrics were averaged across the entire simulation to produce overall performance results. Evaluation was restricted to 13 medical procedures covered by the available dataset and included in the current evaluation. Procedure-event coverage was calculated based on TP detections within the corresponding annotated time intervals.

Detection Effectiveness Measurement

Algorithm performance was evaluated using the defined confusion matrix, in which each algorithm output was classified as a TP, an FP, or an FN. The specific definitions of TP, FP, and FN are detailed in Table 4. Based on these classifications, performance was quantified using 4 standardized metrics as defined in equations 1-4. By structuring the evaluation in this way, the framework ensures that algorithm performance is assessed in a manner consistent with the available annotation granularity, providing meaningful and interpretable metrics that reflect detection effectiveness while accounting for the annotation constraints.

LT Measurement

For each annotated medical procedure within the simulation, LT was measured as the interval from the start of procedure n to the time stamp of the first TP detection (t^n_D) within the corresponding time window. In other words, for each procedure, the earliest TP output was identified, and the LT score was calculated based on this time stamp. The process of identifying the earliest TP for each procedure is illustrated in Figure 5A. The LT score for procedure n was then calculated according to the formula shown in equation 5. The final score ranged from 0 to 1, with earlier detections yielding higher values. The overall simulation-level LT score was obtained by averaging the scores across all procedures.

Medical Object Detection

Overview

The government-generated annotations for the TCCC dataset lacked device-specific temporal granularity, necessitating the establishment of ground truth for medical object detection through indirect methods. Ground-truth medical object labels were derived from the predefined list of devices associated with each procedure, as shown in Table 3. For any annotated procedure, the corresponding device set was designated as the expected set of active devices. Algorithm outputs were compared against these sets, enabling evaluation of detection performance even in the absence of device-specific annotations. Based on this ground truth, medical object coverage was then calculated using the procedure-derived ground truth at the simulation level. As device-specific annotations become available, the active device set can be replaced with device-specific annotations without changing the evaluation framework.

Detection Effectiveness Measurement

Detection effectiveness was evaluated by comparing each algorithm's outputs against the active devices using confusion matrices as illustrated in Figure 5B. All detections were assessed over the full simulation period, enabling assessment even in the absence of temporally resolved annotations. The definitions of confusion matrices are provided in Table 4. Using this framework, performance metrics were calculated in the same manner as for medical procedure detection, following equations 1-4, ensuring a consistent evaluation approach across algorithm types.

LT Measurement

The LT score was computed using the time stamp of the first TP detection relative to the total simulation duration, as defined in the formula shown in equation 5. The process of identifying the earliest TP is illustrated in Figure 5B. Under the current simulation-level evaluation approach, the LT score represents the timing of the first TP detection relative to the total simulation duration rather than direct temporal alignment with object use.

In the current evaluation approach, medical object ground truth was derived from the set of devices associated with each procedure rather than from direct object-use annotations. Therefore, all procedure-associated devices were considered active for evaluation purposes, and detections corresponding to these devices could be counted as TPs even when the object was not explicitly used during the procedure. This represents a limitation of the current evaluation approach because object activity was inferred from procedure annotations rather than direct visual verification of object use. Therefore, Jaccard Index, precision, recall, and F_1 values may be artificially inflated for some objects due to the current ground-truth assumption. These metrics should be interpreted as agreement with the procedure-derived ground truth rather than as direct measures of object detection accuracy. Nevertheless, this approach provides a practical and consistent basis for comparing algorithm performance under the current annotation limitations.

In addition, object detections were evaluated over the full simulation duration rather than at specific time points because temporally resolved object annotations were not available. Consequently, detections made at any point during the simulation could be counted as TPs if they matched the active device set. Although simulation-level evaluation represents an additional limitation of the current approach, it remained consistent with the objective of TCCC documentation, where identifying which medical objects were involved during the overall treatment process was prioritized over determining the precise timing of their use. Therefore, the proposed evaluation framework was designed to assess algorithm outputs against the documented treatment record at the event level rather than to determine the precise temporal localization of individual object detections, reflecting the operational objective of autonomous TCCC documentation.

Injury Detection and Localization

Overview

Injury detection and localization algorithms were evaluated for their ability to find injuries and localize them within predefined body regions (Figure 3), with injury pattern classification excluded from this phase of assessment. When an injury was detected, the algorithm generated a corresponding location, which was mapped to the body regions defined in Figure 3. Entries corresponding to detection failures, such as [], "no detection," NaN, or similar outputs, as well as detections outside the defined body regions, were treated as FNs. Injury coverage was also calculated at the simulation level based on TP detections of annotated anatomical regions.

Detection Effectiveness Measurement

Due to the absence of temporally resolved annotations, algorithm outputs were evaluated across the full simulation period, similar to medical object detection, to ensure comprehensive assessment against the high-level annotation. Confusion matrices were generated by comparing algorithm outputs with the annotated injury locations as illustrated in Figure 5C. Confusion matrix definitions for injury detection and localization are listed in Table 4. Performance metrics were computed in the same manner using equations 1-4.

LT Measurement

The LT score was calculated in the same manner as for medical object detection, using the time stamp of the first TP relative to the total simulation duration. The process of identifying the earliest TP is illustrated in Figure 5C, and the corresponding calculation is formalized in equation 5. As with medical object detection, the LT score represents the timing of the first TP detection.

Similar to medical object detection, simulation-level evaluation represents a limitation of the current approach. Because temporally resolved injury annotations were not available, injury detections were compared against the annotated injury locations at the simulation level rather than within temporally constrained intervals. As a result, temporally inaccurate detections could still be counted as TPs if they matched the annotated injury location. Nevertheless, this evaluation strategy remained consistent with the event-level objective of TCCC documentation and was designed to assess algorithm outputs against the documented treatment record rather than to determine precise temporal localization.

Ethical Considerations

The evaluation team was committed to conducting fair, thorough, and transparent evaluations. To avoid the inadvertent influence of biases on the algorithm evaluation results, a mitigation plan was implemented that included leveraging standardized research protocols, multiple analysts, independent reviewers, documented procedures, calibration verification, and cross-validation of data against expected ranges. The human subject research protocol leveraged to collect the TCCC datasets was reviewed and approved by the United States Army Medical Research and Development Command Institutional Review Board Office, which included human participants' ethics review. The approved protocol (M-11057) was classified as minimal risk to human participants and undergoes annual continuing reviews by the Institutional Review Board Office to ensure the safety and ethical considerations of research on all human participants in our research. Before conducting simulation scenarios, individual participants are provided with a consent form. This form explains the activities, potential risks, and data collection that the principal investigator would like to conduct. If the participant agrees to participate, they are asked to sign the consent form, either in hard copy or electronically. All simulation participants (eg, care providers and simulated patients) are assigned a participant identifier to be associated with collected demographics, video, and physiological signal recordings. All data are stored in Department of War– and Health Insurance Portability and Accountability Act–compliant safety and security compliant e-storage facilities. All hard copy documents including completed consent forms are stored in locked filing cabinets at USAISR. Following data collection sessions, all recordings are checked to ensure that any nonconsented individual's identifiable information was not recorded. If any inadvertent recordings were made, the principal investigator or a study team member deletes

these segments of recordings. All recordings are stored securely in Department of War–approved, USAISR-maintained server or repository including a network-attached storage at Fort Detrick, Maryland. Additionally, photo release waivers for image use in nonresearch activities also accompany the consent forms at the start of participation but are entirely optional for participation in the study. Military volunteers who meet the eligibility requirements and consent to participate in the study had the opportunity to participate in volunteer hours as a participation incentive. After completion of their participation in the study, the participant may request a letter of participation from the research team. The eligibility for volunteer hours toward service recognition will be determined by the units or offices that grant the volunteer hours. The research team will provide the letter of participation in an electronic format to the participant's email address. This process will be mirrored at all data collection locations. There are no monetary incentives or compensation for this research effort for volunteers.

Results

As of September 2025, USAISR completed the performance evaluation of several algorithms across development partners as part of the initial stage of a multiyear program, with additional performance assessments planned over the next 24 months. Because the purpose of this paper is to demonstrate the evaluation methodology rather than to compare algorithm performance, only 1 representative algorithm from each functional category is included. The evaluations cover 3 categories: medical procedure detection, medical object detection, and injury detection and localization. Each algorithm was tested across 15 reserved simulation scenarios. All reported results are drawn exclusively from evaluations conducted in the orchestrated execution environment (ie, phase 2), which reflects the intended operational deployment context.

The results presented here are intended to be interpreted as illustrative examples of the standardized outputs generated by the proposed evaluation framework rather than as definitive measures of algorithm performance or comparative performance assessments. Accordingly, the observations should not be interpreted as generalizable conclusions regarding algorithm performance. Comprehensive results, cross-algorithm comparisons, and detailed performance analyses will be reported separately upon completion of the project. All evaluated algorithms are at an early stage of development and therefore exhibit low performance across the reported metrics, reflecting limited training data and preliminary model optimization. Nevertheless, the evaluation process produces structured and interpretable outputs across all 3 functional categories, demonstrating how the proposed framework systematically organizes performance assessment and establishes a consistent foundation for subsequent integration, comparison, and analysis.

Figure 6 presents the medical procedure detection performance evaluation results. A series of graphical analyses provides an overview of procedure detection performance across the simulation datasets. We also assessed procedure-event coverage across the 15 evaluation simulations. Of the 94 annotated procedure events, 44 had at least 1 TP detection, resulting in an overall coverage of 46.8%, as shown in Table 5.

The F_1 heatmap in Figure 6A shows that detection outcomes vary across procedures and simulations. Procedure-level average F_1-scores ranged from 0.00 to 0.31, while simulation-level averages ranged from 0.00 to 0.33. Nasopharyngeal airway, chest seal application, and treat a casualty for a cold injury show higher F_1-scores (mean F_1=0.28, 0.31, and 0.31, respectively), whereas several other procedures show near-zero or low F_1-scores.

The boxplots in Figure 6B show procedure-specific average F_1-scores across all simulations and highlight performance variability across procedures, with several procedures exhibiting wide interquartile ranges and low medians. Importantly, these visualizations illustrate observable patterns that may reflect data and model limitations by identifying procedures for which performance is consistently low or highly variable, thereby suggesting areas where additional training data and focused model refinement may be beneficial. The results serve not only as performance summaries but also as example outputs of the evaluation framework for prioritizing data collection and targeted retraining efforts for subsequent algorithm development.

The radar chart in Figure 6C summarizes average performance metrics across all simulations, with output-level Recall showing relatively higher values than the other metrics. While overall performance is low, these analyses illustrate the types of results produced through the assessment pipeline and highlight areas where further refinement and larger training datasets will be essential.

LT scores varied across simulations, with most simulations with available LT measurements achieving scores between approximately 0.6 and 0.9. The average LT score was 0.735 (11 of 15 simulations), reflecting the timing of the first TP detection for corresponding procedures across the simulations. LT scores were not available for simulations in which the algorithm failed to detect any TPs. Together with the detection performance metrics, these results provide an additional assessment of the temporal characteristics of the representative procedure detection algorithm.

Figure 6. Medical procedure detection performance evaluation results. (A) Procedure-specific F_1-scores across simulations; (B) distributions of F_1-scores for each procedure across simulations; and (C) average precision, recall, F_1, and Jaccard Index. IV: intravenous.

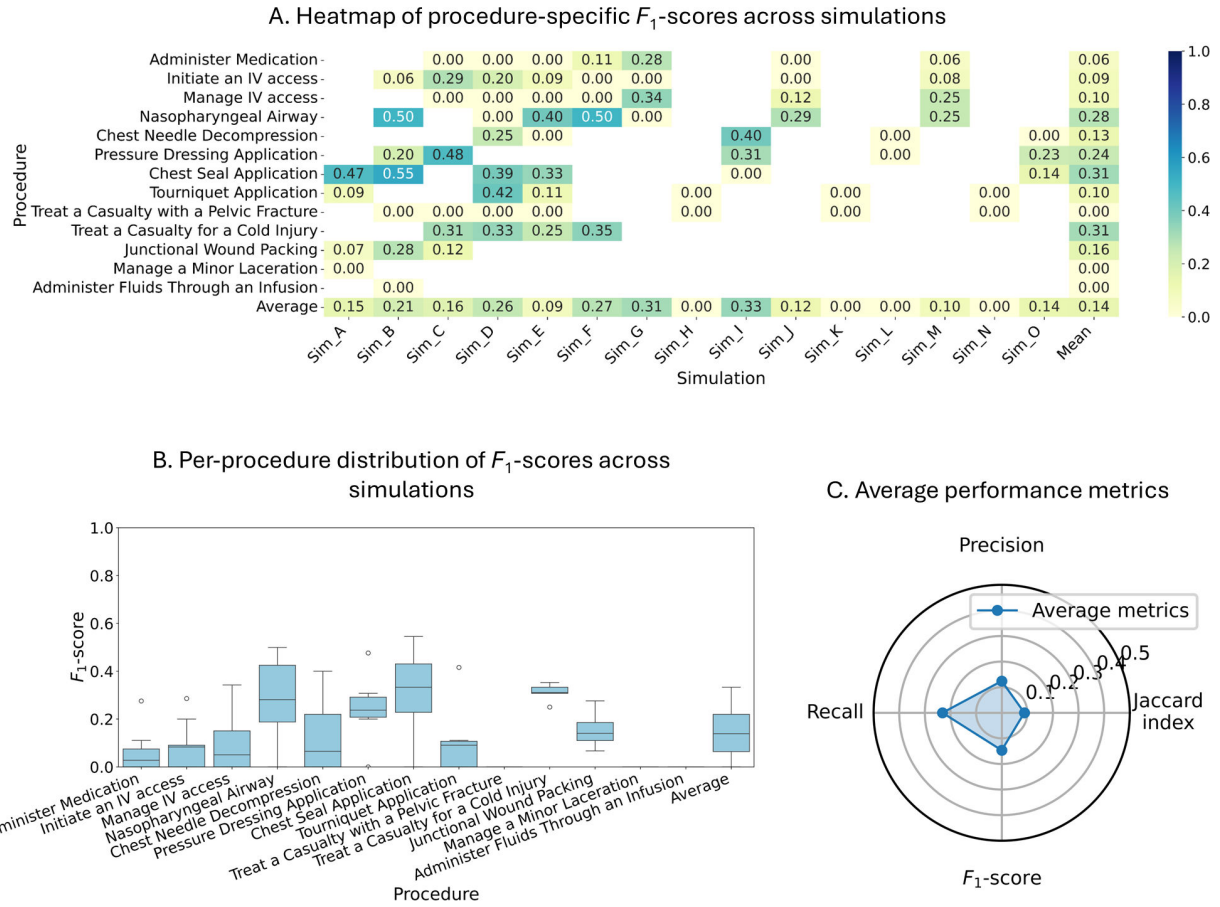


Table 5. Ground-truth coverage across the 15 evaluation simulations[a].

Simulation	Medical procedure detection, n/N (%)	Medical object detection, n/N (%)	Injury detection and localization, n/N (%)
Sim A	4/9 (44.4)	4/10 (40)	3/5 (60)
Sim B	5/7 (71.4)	8/19 (42.1)	3/5 (60)
Sim C	4/9 (44.4)	8/13 (61.5)	2/4 (50)
Sim D	8/14 (57.1)	12/17 (70.6)	4/6 (66.7)
Sim E	5/12 (41.7)	10/17 (58.8)	4/6 (66.7)
Sim F	3/7 (42.9)	2/9 (22.2)	2/5 (40)
Sim G	5/8 (62.5)	2/8 (25)	0/4 (0.0)
Sim H	0/2 (0.0)	0/2 (0.0)	0/3 (0.0)
Sim I	2/3 (66.7)	4/9 (44.4)	4/4 (100)
Sim J	2/7 (28.6)	6/8 (75)	2/5 (40)
Sim K	0/2 (0.0)	2/2 (100)	3/5 (60)
Sim L	0/2 (0.0)	5/5 (100)	2/2 (100)
Sim M	4/6 (66.7)	7/8 (87.5)	4/4 (100)
Sim N	0/2 (0.0)	2/2 (100)	2/4 (50)
Sim O	2/4 (50)	6/9 (66.7)	3/3 (100)
Overall	44/94 (46.8)	78/138 (56.5)	38/65 (58.5)

[a]Values represent the number of annotated items with at least 1 true positive detection divided by the total number of annotated items. Procedure-event coverage was assessed at the event level, while medical object and injury coverage were assessed at the simulation level.

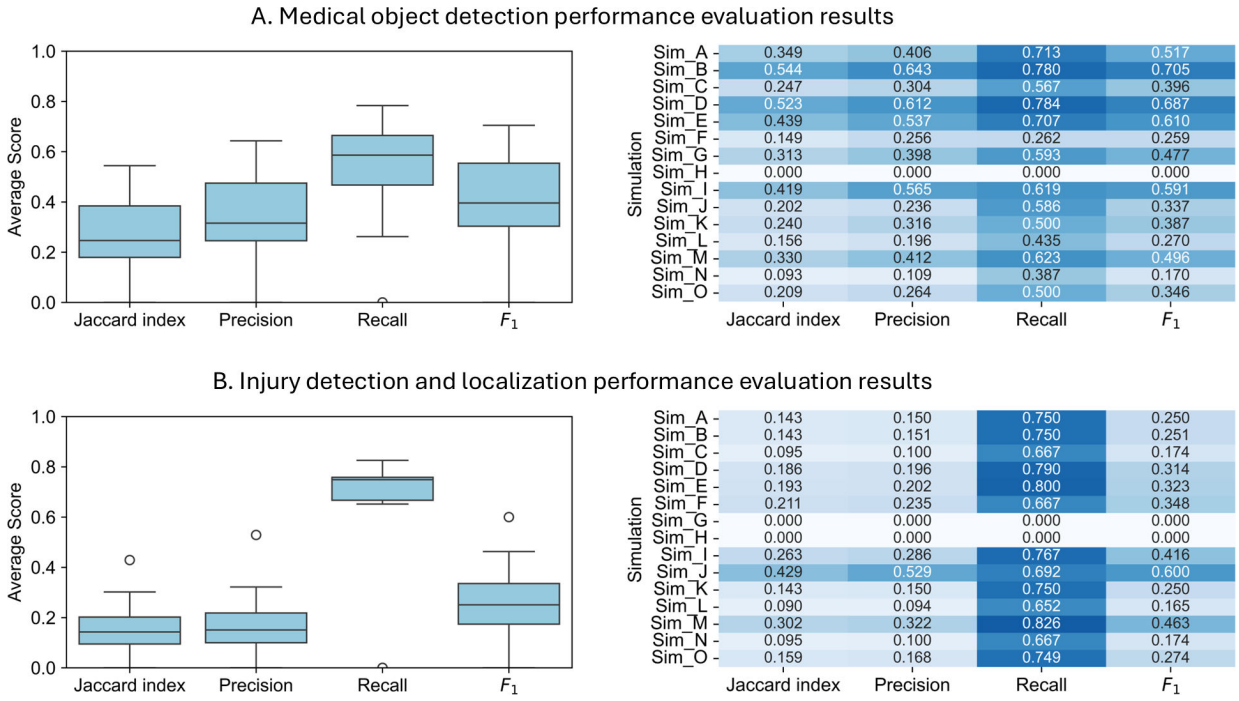
Figure 7 summarizes the performance evaluation results for medical object detection and injury detection and localization. Figure 7A shows the medical object detection results, whereas Figure 7B shows the injury detection and localization results.

The boxplot in the left panel of Figure 7A shows the distribution of medical object detection performance metrics,

with relatively higher output-level Recall values, although overall results remain limited. The heatmap in the right panel of Figure 7A highlights simulation-specific variability, with certain cases such as Sim B and Sim D performing comparatively better across all 4 metrics, although these performances are not consistent across other simulations. F_1-scores ranged from 0.000 to 0.705, with performance varying substantially across scenarios. The simulation-specific variability also provides useful guidance for algorithm developers by suggesting simulation conditions, such as lighting, environment, and patient presentation, that may require additional training data and further model refinement. For medical object coverage, 78 of the 138 medical objects had at least 1 TP detection across the 15 evaluation simulations, resulting in an overall coverage of 56.5%, as shown in Table 5. These results illustrate both the potential and the current limitations of the detection pipeline, while reinforcing the importance of continued development.

The boxplot in the left panel of Figure 7B shows the distribution of injury detection and localization performance metrics, with output-level Recall relatively higher than the other metrics. The heatmap in the right panel of Figure 7B shows performance differences across simulations, with some instances such as Sim J and Sim M showing relatively stronger performance, although performance overall remains low. F_1-scores ranged from 0.000 to 0.600, with performance varying substantially across scenarios. These simulation-level differences also provide useful guidance for targeted data collection and further model refinement under various simulation conditions. For injury detection and localization, ground-truth coverage was 58.5%, with at least 1 TP detection for 38 of the 65 annotated anatomical regions across the 15 evaluation simulations, as shown in Table 5.

Figure 7. Medical object detection and injury detection and localization performance evaluation results. (A) Medical object detection performance summarized by metric distributions and simulation-level results; and (B) injury detection and localization performance summarized by metric distributions and simulation-level results. Sim: simulation.



For medical object detection, LT scores were consistently high among simulations with available LT measurements, with most simulations achieving scores above 0.8. The average LT score was 0.881 (14 of 15 simulations). Under the current simulation-level evaluation approach, this score represents the timing of the first TP detection relative to the total simulation duration rather than direct temporal alignment with object use.

Similar to the medical object detection results, the injury detection and localization LT scores were high among simulations with available LT measurements, with an overall average score of 0.912 (13 of 15 simulations). As with medical object detection, the LT score represents the timing of the first TP detection under the current approach. The ground-truth coverage results for all 3 functional categories across the 15 evaluation simulations are summarized in Table 5.

Overall, these outcomes illustrate the role of the evaluation framework in producing standardized results across 3 core algorithm functions. In addition to the detection effectiveness, the incorporated LT evaluation provides a complementary assessment of processing efficiency. The ground-truth coverage analysis provides additional information on the detection of annotated items that is not fully captured by output-level Recall, helping identify potential gaps in detection effectiveness. While algorithm performance remains low, the framework captures both variability and recurring patterns across simulations, reflecting strengths,

limitations, and areas requiring refinement. Accordingly, the performance results in this section are intended to illustrate the operational usefulness of the evaluation workflow for systematic and reproducible algorithm assessment. In addition to algorithm-level performance, system-level resource usage was evaluated to characterize the computational efficiency of algorithm execution under the orchestration framework on the edge-computing device. A systematic evaluation quantified memory consumption as well as central processing unit (CPU) and graphics processing unit (GPU) usage during concurrent execution of all integrated algorithms and models. Dedicated software tools and scripts were developed to automate resource monitoring and generate standardized performance graphs for reproducible system-level analysis.

Since the system is designed to execute multiple algorithms simultaneously on a single resource-constrained edge device, a system-level resource usage benchmark was performed to evaluate the orchestration framework under concurrent execution. The algorithm labels (A-D) shown in Figure S1A and B in Multimedia Appendix 1 represent the algorithms successfully integrated from 4 independent development partners to date. Because algorithm B provided both the medical object detection and medical procedure detection pipelines, 5 functional pipelines were evaluated. The benchmark included 1 injury detection and localization pipeline, Injury (A), 1 medical object detection pipeline, Object (B), 1 medical procedure detection pipeline, Procedure (B), and 2 additional object detection pipelines, Object (C) and Object (D). The corresponding system-level benchmark results are presented in Figure S1A and B in Multimedia Appendix 1.

The memory usage results presented in Figure S1A in Multimedia Appendix 1 summarize the memory footprint of multiple algorithms integrated into the orchestration framework. In this benchmark, heterogeneous algorithms were codeployed and concurrently executed on a single edge device through a unified orchestration layer, which manages both inference services and downstream processing pipelines. The results highlight the memory consumption attributable to distinct system components, including the messaging layer (ZMQ), Triton inference backends (CPU and GPU), and individual application pipelines. While the ZMQ broker exhibits stable and low memory usage across all algorithms, noticeable variability is observed in the Triton CPU and GPU components, reflecting differences in model architectures, batch handling, and back-end execution paths. Algorithm C demonstrates a substantially higher GPU memory footprint, indicating a more resource-intensive inference workload than other algorithms. Memory usage associated with individual application pipelines also varies across functions, reflecting differences in pipeline complexity, preprocessing requirements, and intermediate buffering strategies.

The resource usage results presented in Figure S1B in Multimedia Appendix 1 complement the memory analysis by reporting average resource usage percentages for the same system configuration. Overall usage patterns reveal that the orchestration framework efficiently isolates algorithm-level workloads, preventing excessive contention at the ZMQ layer. CPU and GPU usage profiles further emphasize algorithm-dependent behavior, with Algorithm C again exhibiting higher CPU and GPU usage, consistent with its elevated memory demand. Pipeline-level usage measurements indicate that downstream processing stages, such as the injury detection, object detection, and procedure detection pipelines, contribute nonnegligible but bounded overheads relative to core inference execution. Collectively, these results indicate that the proposed orchestration framework is capable of hosting multiple algorithms on a single edge device while maintaining predictable and analyzable resource consumption, which may support system-level capacity planning and deployment decisions.

Discussion

This study presents a standardized evaluation framework for assessing the performance of algorithms in an autonomous TCCC patient documentation system under both algorithm-level and orchestration-aware operational conditions. This approach enables reproducible and structured performance evaluation of heterogeneous algorithms across multiple simulation scenarios. The framework further demonstrates the feasibility of evaluating algorithm behavior under standardized edge-computing resource constraints representative of operational deployment environments and establishes an initial benchmark for algorithmic capabilities in automatically populating the TCCC Card under realistic operational conditions. In addition, the framework enables reproducible, cross-dataset comparisons, allowing evaluators to quantify algorithm performance using unified metrics and event-level ground-truth annotations. Such standardization is particularly valuable given the diversity of algorithmic approaches contributed by multiple algorithm developers, providing a common basis for performance evaluation as these algorithms continue to mature.

Application of this framework provides critical insights into algorithmic performance across diverse functionalities and across the iterative development cycle. The inference models developed to date are inherently limited due to the short 6-month development cycle and the fact that they were trained on only roughly a quarter (75 unique TCCC simulations) of the total data collected to date. Additionally, the 15 simulations reserved for algorithm evaluation include a higher degree of diversity of conditions that were either underrepresented in the training data or completely missing, such as low lighting conditions and red-light illumination. Given these limitations of the training data, the low performance of the inference models in their current state is expected. The scale and diversity of data available for training will increase over the next 24 months as more data become available, resulting in a larger training dataset that more closely matches the evaluation dataset held in reserve. Therefore, the inference model performance is expected to increase in subsequent algorithm training and development cycles.

Although the evaluation results presented in this phase are based on early-stage inference models, the framework's primary value lies in revealing each detection

model's strengths and weaknesses. By leveraging both simulation-level evaluations and detection-specific performance aggregated across simulations, the framework enables targeted identification of underperforming areas, supporting iterative and strategic data collection and model refinement. By providing reproducible metrics and structured evaluation, it not only benchmarks current capabilities but also establishes a foundation for continuous, evidence-based advancement in algorithm development.

This study has several limitations. While the evaluation framework offers valuable insights into algorithm performance through its standardized workflow, this phase of evaluation was necessarily limited to baseline core functionalities due to the early stage of system development. Given the focus on limited annotations and features commonly implemented across all algorithms in this phase, certain capabilities such as injury pattern recognition, scenarios involving multiple patients, and the use of physiological inputs such as vital sign monitors and electromyography armbands could not be included. The assessment therefore concentrated on a single patient and relied solely on video-based data. Evaluations also focused primarily on individual algorithm outputs and module-level characterization, rather than combined outputs from multiple algorithms. These limitations highlight opportunities to extend the framework to more complex, multimodal, and system-level scenarios, supporting comprehensive, generalizable, and operationally relevant performance benchmarks.

Another limitation is the lack of higher-granularity annotations. As a result, the current ground truth definitions and simulation-level evaluation were adopted where appropriate under the event-level documentation objective. In addition, the current Recall metric is calculated at the algorithm-output level and does not measure annotated item coverage. Ground-truth coverage was therefore assessed separately using the available annotations. These coverage measures are limited by the available annotation granularity. In particular, simulation-level coverage for medical object detection and injury detection and localization does not distinguish repeated occurrences of the same object class or anatomical region within a simulation or assess the temporal accuracy of individual detections. These limitations will be addressed as higher-granularity annotations become available in future datasets. Task-level temporal overlap was assessed as part of the annotation quality control. However, this assessment focused on annotation timing and did not measure agreement on task identification separately. In addition, an interannotator reliability estimate was not calculated for the current dataset, which represents a limitation of the current annotation quality assessment.

At the current stage, the algorithm orchestration framework is designed to support parallel execution of multiple algorithms under edge-computing constraints, where system resources, timing, and synchronization are inherently limited. However, for algorithm performance evaluation, each algorithm was evaluated individually, whereas the system-level resource usage benchmark was conducted with multiple integrated algorithms executed concurrently. This study prioritizes individual algorithm (module-level) evaluation to isolate and more accurately characterize the performance of each algorithm within a workflow-aware and orchestration-enabled assessment framework before full system integration and validation. Full integration of the consolidated outputs across algorithms will be addressed in subsequent stages. This evaluation scope enables clearer identification of performance bottlenecks, latency sources, and resource constraints that may not be observable in fully integrated outputs.

Ongoing data collection continues to expand the dataset pool, with all data systematically processed through a government laboratory pipeline for classification, annotation, and clearance. Acquiring higher-granularity, evaluation-ready datasets with more precise annotations will be critical to overcome current limitations and enhance assessment fidelity. Techniques such as digital twinning can provide the temporal and spatial resolution necessary for rigorous algorithm training and evaluation, thereby supporting reliable real-world deployment. Following each round of model updates, the standardized evaluation framework will be applied across the expanded datasets, enabling large-scale assessments with improved statistical rigor and comparative analyses across algorithms.

While this study was developed in the context of military TCCC patient documentation, the proposed evaluation framework is not limited to this specific operational setting. The core design principles of this framework, including event-level evaluation, standardized performance metrics, and structured comparison across different algorithms, are broadly applicable to other clinical and digital health domains. By decoupling the evaluation methodology from specific TCCC procedures, the framework provides a generalizable structure for assessing algorithm performance in diverse real-world clinical environments such as hospital-based clinical documentation systems, emergency medical services, and other digital health care applications. As such, the proposed framework has potential relevance beyond military use, supporting broader adoption in clinical AI systems that require consistent, interpretable, and reproducible performance evaluation.

Future assessments will extend to multiple patient scenarios and mass casualty environments, introducing additional operational complexity and new variables that challenge current evaluation methodologies. These assessments will address tasks such as accurately distinguishing individual patients, identifying objects across repeated and distinct occurrences, handling variable procedure sequences, and managing dynamic environmental conditions with multihuman interactions. Future versions of the framework will also support concurrent and overlapping medical procedures, enabling evaluation under more realistic clinical workflows in which multiple interventions occur simultaneously. By systematically evaluating performance under these conditions, the study aims to measure algorithmic robustness, generalizability, and system-level reliability. The resulting insights will guide iterative improvements to algorithmic architectures and targeted expansion of training datasets, supporting future integrated autonomous TCCC

patient documentation systems in diverse and unpredictable environments.

In conclusion, the proposed framework establishes an initial foundation for standardized and operationally relevant evaluation of autonomous patient documentation algorithms in TCCC environments. Beyond benchmarking current algorithm capabilities, the framework provides a scalable structure for future multimodal, orchestration-aware, and clinically integrated AI systems. As algorithm maturity, dataset diversity, and system integration continue to improve, the framework may support broader development of reliable and interpretable clinical AI technologies for both military and civilian health care applications.

Acknowledgments

This autonomous tactical combat casualty care documentation algorithm assessment was a collaborative, USAISR-wide effort led by members of the Medical Robotics and Autonomous Systems and Data Science Division teams. Additionally, the USAISR team would like to recognize the contributions of the Johns Hopkins Applied Physics Lab (JHU APL) team for their support with the algorithm orchestration assessment effort. The following individuals are acknowledged for their direct and indirect support to this effort: Dr Shawn Burgener, Dr Jonathan Chambers, Dr Brant Chee, Mr James Gaudaen, Mr Peter Chung, Mr Valmik Desai, Ms Alix Donnelly, Mr David Drewry III, Ms Stephanie Edwards, Mr Tyler Feldman, Dr Amanda Galante, Dr Prudhvi Gurram, Ms Tina Hamrick, Mr Wesley Huff, Mr Jakob Kaminsky, Ms Buffy Krall, Dr Yemeserach Mekonnen, LTC (Ret) Rachel Morgans, COL (Ret) Jeremy Pamplin, Mr Matt Quinn, Mr Jeff Robbins, Ms Kate Tallaksen, Ms Ha Tran, and SFC Justin Warren.

The views expressed are those of the authors and do not reflect the official policy of the Department of the Army, the Department of War, the Defense Health Agency, or the US federal government.

Generative AI Use

The authors did not use generative AI tools to develop this manuscript. However, during the preparation of this work, the authors used the Grammarly tool to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

The funding for this project (#CO230044) was provided under the Enabling Medical System Technologies (EMST) Program Area, Combat Casualty Care (CCC) Portfolio, Defense Health Agency, Research and Engineering. The EMST CCC portfolio was previously referred to as the Joint Program Office 6 Combat Casualty Care Research Program (JPC-6 CCCRP). The funder did not influence the design of the study or strategies related to its collection, analysis, or interpretation of data.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to limited use restrictions of our existing research protocol(s) and the government requirement for data use agreements for new partnerships.

Authors' Contributions

J Lee and J Little drafted the original manuscript. J Lee, EQ, and NF designed this research assessment approach. J Lee, J Little, EQ, OB, and NF edited the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Resource usage benchmark results of the edge orchestration framework. (A) Memory usage and (B) resource usage of the ZMQ broker, Triton CPU and GPU components, and individual pipelines during concurrent execution on a single edge-computing device. CPU: central processing unit; GPU: graphics processing unit; ZMQ: ZeroMQ.
[PNG File (Portable Network Graphics File), 193 KB-Multimedia Appendix 1]

References

1. Tactical combat casualty care (TCCC) guidelines for all service members. Joint Trauma System. Jan 25, 2024. URL: https://jts.health.mil/index.cfm/committees/cotccc/guidelines [Accessed 2026-02-10]
2. Army Futures Command Concept for Medical 2028, AFC Pamphlet 71-20-12 US Army Medical Center of Excellence. US Army Futures Command. 2022. URL: https://api.army.mil/e2/c/downloads/2022/04/25/ac4ef855/medical-concept-2028-final-unclas.pdf [Accessed 2026-02-10]
3. Patel J, Boardman M, Files B, et al. Give us a hand, mate! A holistic review of research on human-machine teaming. BMJ Mil Health. Sep 22, 2025;171(5):392-396. [doi: 10.1136/military-2024-002737] [Medline: 39719381]
4. Leone RM, Chambers JA, Donham BP, Junker CA. Artificial intelligence in military medicine. Mil Med. Aug 30, 2024;189(9-10):244-248. [doi: 10.1093/milmed/usae359]

5. Veigl A, Wardall A, Butkus C. Smart device evaluation for deployed operations: final report. 711th Human Performance Wing, Airman Systems Directorate, Air Force Research Laboratory, United States Air Force; 2021. URL: https://apps.dtic.mil/sti/pdfs/AD1152865.pdf [Accessed 2026-02-09]
6. Reducing the cognitive burden on the modern soldier: leveraging SlingShot and GENSS technologies. Asian Military Review. Dec 16, 2024. URL: https://www.asianmilitaryreview.com/2024/12/reducing-the-cognitive-burden-on-the-modern-soldier-leveraging-slingshot-and-genss-technologies/ [Accessed 2026-02-04]
7. Kodithuwakku Arachchige SNK, Burch RFV, Chander H, Turner AJ, Knight AC. The use of wearable devices in cognitive fatigue: current trends and future intentions. Theor Issues Ergon Sci. May 4, 2022;23(3):374-386. [doi: 10.1080/1463922X.2021.1965670]
8. Gaudaen JC, Papadopoulos A, Dockery T, Manemeit C. Usability enhancements to a prototype clinical decision support system for combat medics. Mil Med. Nov 8, 2023;188(Suppl 6):614-620. [doi: 10.1093/milmed/usad279] [Medline: 37948290]
9. Lutz J, Memmert D, Raabe D, Dornberger R, Donath L. Wearables for integrative performance and tactic analyses: opportunities, challenges, and future directions. Int J Environ Res Public Health. Dec 19, 2019;17(1):59. [doi: 10.3390/ijerph17010059] [Medline: 31861754]
10. Hunter OF, Perry F, Salehi M, et al. Science fiction or clinical reality: a review of the applications of artificial intelligence along the continuum of trauma care. World J Emerg Surg. Mar 6, 2023;18(1):16. [doi: 10.1186/s13017-022-00469-1] [Medline: 36879293]
11. Alshamari MA, Althobaiti MM. Usability evaluation of wearable smartwatches using customized heuristics and System Usability Scale Score. Future Internet. 2024;16(6):204. [doi: 10.3390/fi16060204]
12. Mejia C, Ciarlante K, Chheda K. A wearable technology solution and research agenda for housekeeper safety and health. Int J Contemp Hosp Manag. Oct 20, 2021;33(10):3223-3255. [doi: 10.1108/IJCHM-01-2021-0102]
13. Shackel B, Richardson SJ. Human Factors for Informatics Usability. E-book, Google Books: Cambridge University Press; URL: https://books.google.com.mt/books?id=KSHrPgLlMJIC&printsec=copyright&source=gbs_pub_info_r#v=onepage&q&f=false [Accessed 2026-02-04]
14. Stoor-Burning E, Gaudaen J, Pavliscsak H, Little J. Candidate passive sensor suite technologies for tactical combat casualty care (TCCC) environments: results of a systematic, multi-phased comparative assessment. JMIR Preprints. Preprint posted online on Sep 26, 2025. [doi: 10.2196/preprints.84900]
15. Little JR, Rivera-Nichols T, Pavliscsak HH, et al. Demonstrating tactical combat casualty care in simulated environments to enable passive, autonomous documentation: protocol for a prospective simulation-based study. JMIR Res Protoc. Mar 17, 2025;14:e67673. [doi: 10.2196/67673]
16. Schmidt PM, Ortman H, Gaudaen JC, et al. Developing a comparative effective methodology for technology usability during a simulated casualty event. Mil Med. Nov 8, 2023;188(Suppl 6):642-650. [doi: 10.1093/milmed/usad289] [Medline: 37948220]
17. Pamplin JC, Remondelli MH, Thota D, et al. Revolutionizing combat casualty care: the power of digital twins in optimizing casualty care through passive data collection. Mil Med. Jan 16, 2025;190(1-2):27-32. [doi: 10.1093/milmed/usae249] [Medline: 38743585]
18. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. Oct 29, 2019;17(1):195. [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
19. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data. Jun 17, 2019;6(1):96. [doi: 10.1038/s41597-019-0103-9] [Medline: 31209213]
20. Payne TH, Graham G. Managing the life cycle of electronic clinical documents. J Am Med Inform Assoc. 2006;13(4):438-445. [doi: 10.1197/jamia.M1988] [Medline: 16622169]
21. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. Jan 1, 2013;20(1):117-121. [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]

Abbreviations

CPU: central processing unit
FN: false negative
FP: false positive
GPU: graphics processing unit
HMT: human-machine teaming
LT: lag time
TCCC: tactical combat casualty care
TP: true positive
USAISR: United States Army Institute of Surgical Research

ZMQ: ZeroMQ

Edited by Ivan Steenstra; peer-reviewed by Holly H Pavliscsak, Mehmet Tatli; submitted 25.Mar.2026; initial comments to author 01.May.2026; final revised version received 22.Oct.2026; accepted 22.Oct.2026; published 24.Nov.2026

Please cite as:
Lee J, Little J, Quist E, Badawi O, Fisher N
Standardized Performance Assessment Methodology and Evaluation Framework for Tactical Combat Casualty Care Autonomous Documentation Algorithms: Framework Development and Evaluation
JMIR Med Inform 2026;14:e95075
URL: https://medinform.jmir.org/2026/1/e95075
doi: 10.2196/95075

© Jaeyeon Lee, Jeanette Little, Ethan Quist, Omar Badawi, Nathan Fisher. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 24.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.