Data Competition Rules

Challenge Event 2

Version 2d

June 20, 2025

Defense Advanced Research Projects Agency
Biological Technologies Office
675 North Randolph Street
Arlington, VA 22203-2114
TriageChallenge@darpa.mil

1 Contents

2 Introduction

This document describes the Data Competition Rules of the DARPA Triage Challenge (DTC). This document supersedes previous versions of the DARPA Triage Challenge Rules. Significant revisions from past versions in this document are indicated by blue text. Teams are encouraged to closely review the entire document. The intent of this document is to provide participants guidance on competition design and scoring objectives to inform their development efforts in preparation for the first competition event. This document is subject to change and may be superseded by later versions. The latest official versions of all documents are posted on the DARPA Triage Challenge Website (triagechallenge.darpa.mil) and the DARPA Triage Challenge Community Forum.

DARPA intends to release a draft of the Competition Rules no later than nine months before each Challenge Event. The final version of the Competition Rules will be released no later than three months prior to each respective event.

The DARPA Triage Challenge Chief Judge has the final authority to make any decisions related to the rules or scoring. All decisions made by the Chief Judge are final.

The main goal of the DARPA Triage Challenge is to inspire development of scalable, timely, and accurate capture of novel injury signatures to enhance triage decision-making in austere, complex, and mass-casualty settings. The challenge elements and the competition structure itself are intended to address the additional goal of increasing the diversity, versatility, cost-effectiveness, and robustness of relevant technologies and systems capable of addressing the myriad needs of a wide range of mass casualty incidents (MCIs) rather than single-purpose or specifically tailored solutions. The third goal of the DARPA Triage Challenge is to establish a collaborative community by bringing together multi-disciplinary teams and cross-cutting approaches across disparate fields to address the autonomy, perception, and diagnostic needs of the medical triage community.

3 Overview

Under the authority of 10 U.S.C. §4025 to stimulate innovations using prize competition, the DARPA Triage Challenge will use a series of competition events to drive breakthrough innovations in the identification of physiological features ("signatures") of injury. These new signatures will help medical responders perform scalable, timely, and accurate triage. Of particular interest are MCIs, in both civilian and military settings, when medical resources are limited relative to the need.

The DARPA Triage Challenge's long-term vision is 1) an initial, or primary stage of MCI triage supported by sensors on stand-off platforms, such as uncrewed aircraft vehicles (UAVs) or uncrewed ground vehicles (UGVs), and algorithms that analyze sensor data in real-time to identify casualties for urgent hands-on evaluation by medical personnel; followed by 2) a secondary stage, after the most urgent casualties have been treated, supported by non-invasive sensors placed on casualties and algorithms that analyze sensor data in real-time to predict the need for life-saving interventions (LSIs) by medical personnel. Injury information provided by these sensors in primary and secondary triage could be integrated with other information about the scene to accumulate evidence about the injury mechanism and characteristics in order to enhance overall situational awareness, and to focus further physiological interventions.

To advance progress towards this vision, the DARPA Triage Challenge aims to bring together multi-disciplinary teams and industries that will identify physiological signatures and develop sensor and algorithm strategies for complex MCI settings. Teams participating in the DARPA Triage Challenge will be

[1] Patterns in sensor data that reflect or predict injuries of high importance for triage assessments

tasked with developing and demonstrating strategies for capturing high-value signatures for either primary or secondary triage, or for both. While aspects of the DARPA Triage Challenge involve sensors and sensor-delivery platforms, the priority is the development of physiological signatures and models to detect them, not the development of new sensor or platform technology.

4 DARPA Triage Challenge Schedule Overview

The DARPA Triage Challenge is a 3-year effort with 3 sequential 12-month phases for Primary Triage (Systems Competition) and Secondary Triage (Data Competition) in parallel, each culminating in a challenge event (Figure 1; see the DTC website for competition details). In each phase, competitors will develop signatures and detection and analysis strategies for Primary and/or Secondary Triage. DARPA will host two competition events in each phase; a workshop and a challenge event.

Competition events will become progressively more difficult and realistic from Phase 1 to Phase 3.

The workshops will provide an opportunity for practice and mid-phase evaluation for all tracks.

Table 1 provides additional information on schedule and format of Competition events and workshops.

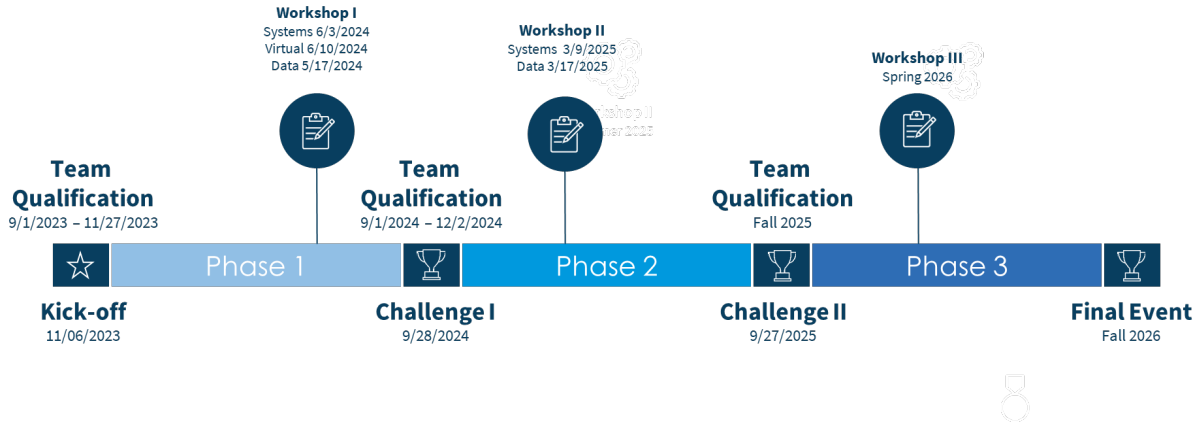

Figure 1 - Program structure and schedule for the DTC.

Data Competition - Tracks D and E			
Event	**Location**	**Est. Duration**	**Date**
Year 1			
Challenge Kick-off	Hybrid	2 days	Nov 6-7, 2023
Workshop - Month 8 *Evaluations / Runs*	Virtual	7 days	May 17, 2024 (Data)
Workshop - Month 8 *Lessons-learned session*	Virtual	1 day	6/17/2024
Challenge 1 - Month 12 *Evaluations / runs*	Virtual	TBD	8/30/2024
Challenge 1 - Month 12 *Awards /lessons-learned session*	Hybrid	1 day	10/5/2024
Year 2			
Workshop - Month 4 *Evaluations / Runs*	Virtual	7 day	3/17/2025
Workshop - Month 4 *Lessons-learned session*	Virtual	1 day	Spring 2025
Challenge 2 - Month 12 *Evaluations / runs*	Virtual	TBD	8/30/2025
Challenge 2 - Month 12 *Awards /lessons-learned session*	Hybrid	1 day	10/04/2025
Year 3			
Workshop - Month 4 *Evaluations / Runs*	Virtual	7 day	Spring 2026
Workshop - Month 4 *Lessons-learned session*	Virtual	1 day	Spring 2026
Final Challenge - Month 11 *Preliminary rounds*	Virtual	TBD	Fall 2026
Final Challenge - Month 11 *Finalists only - Runs and Awards*	In person	1 day	Fall 2026

*Table 1 - Schedule of DARPA-organized Challenge events and workshops. *Note: DARPA-funded teams must participate in all workshop events. It is highly recommended that self-funded Systems teams also attend the workshops.*

5 Prizes and Funding

Teams are encouraged to pursue high-risk, high-reward approaches to meet and exceed the objectives of the Challenge Events. Monetary prizes will be awarded for the Systems and Data Competitions at each of the Challenge Events (Table 2).

Challenge I Fall 2024

Systems [self-funded]	Data [self-funded]	Virtual [self-funded]
1st $120,000	1st $120,000	1st $60,000
2nd $60,000	2nd $60,000	2nd $30,000
3rd $20,000	3rd $20,000	3rd $10,000

Challenge II Fall 2025

Systems [self-funded]	Data [self-funded]
1st $300,000	1st $300,000
2nd $150,000	2nd $150,000
3rd $50,000	3rd $50,000

FINALS Fall 2026

Systems [DARPA-Funded and self-funded]	Data [DARPA-Funded and self-funded]
1st $1,500,000	1st $900,000
2nd $750,000	2nd $450,000
3rd $250,000	3rd $150,000

Table 2 - Prize structure for the three Challenge Events.

DARPA-Funded Teams

DARPA-funded teams (Systems and Data Competitions) are only eligible for the prizes in the Final Events (selection for DARPA-funded teams has closed). The Government's obligation for prizes under DTC is subject to the availability of appropriated funds from which payment for prize purposes can be made. No legal liability on the part of the Government for any payment of prizes may arise unless appropriated funds are available to DARPA for such purposes.

Self-Funded Teams

Self-funded teams in the Data Competition are eligible for prizes in all of the Challenge Events.

Data Competition Prizes and Funding: Phase 2 prizes for self-funded teams will be awarded to the best performing self-funded Data Teams, provided that the teams exceed minimum benchmark standards (see Section 9.6.9). High-performing teams are also eligible to become a DARPA-funded team in Phase 3. The Government's obligation for prizes under DARPA Triage Challenge is subject to the availability of appropriated funds from which payment for prize purposes can be made. No legal liability on the part of the Government for any payment of prizes may arise unless appropriated funds are available to DARPA for such purposes.

To be eligible for prizes, teams must first be registered in the team qualification portal. The award process requires recipients to furnish information that may trace or identify recipients either individually or as an organization (e.g., Social Security Number or Tax Identification Number). The primary contact of each registered team is responsible for providing the award information necessary for prize disbursement. DARPA will reach out by email to the primary contact of each registered team to either confirm their vendor status or request the required forms (e.g., SF-3881 or PIF). DARPA is not responsible for disbursement of prizes to any team members other than the primary contact/organization.

At the end of each competition event, teams will be invited to discuss their technical approaches and lessons learned in a townhall-style hotwash. The extent of technical details shared does not need to exceed data agreements established upon qualification.

6 Qualifications

Prospective DTC competitors must demonstrate track-appropriate performance capabilities to be eligible to participate in DARPA Triage Challenge. All teams in all competitions (Primary Triage Systems tracks and Secondary Triage Data tracks; see the DTC website for track details) must complete two types of qualification: a Team Qualification at the beginning of each phase, followed by event-specific Event Qualifications for each Workshop and Challenge Event. Successful Team Qualification is a prerequisite to Event Qualifications in the same phase.

The initial *DTC Event Qualification Guide* is expected to be released by February 18th, 2024. The *DTC Event Qualification Guide* will continue to be updated for each event. The latest revision will be posted on the DTC Website and DTC Community Forum.

6.1 Team Qualification

Teams must qualify for DARPA Triage Challenge competition events during the designated qualification window by completing the *Team Qualification* form on the DTC Team Portal. Team Qualification submissions will be accepted on a rolling basis but must be submitted by the deadline (see Table 3). Team qualification is required to receive access to datasets and must be completed prior to event-specific enrollment.

Team Qualification Windows by Phase	
Phase 1	9/1/2023 - 11/13/2023
Phase 2	9/1/2024 - 12/2/2024
Phase 3	Fall 2025

Table3. – Team qualification schedule.

6.2 Event Qualification

Prospective teams are required to demonstrate baseline performance and utility capabilities (e.g., safety measures for the Systems Competition and algorithm capability for the Data Competition), to be eligible to participate in events. **All** teams (DARPA-funded and self-funded) in all competitions (Systems and Data) must qualify for each event including the DTC workshops, Preliminary Events (i.e. Phase 1 and Phase 2 Challenge Events), and Final Event.

The latest revision of the *DTC Event Qualification Guide* will be posted on the DARPA Triage Challenge Website and DTC Discourse Community Forum. Event Qualification submissions will be accepted on a rolling basis but must be submitted by the deadline to be eligible to participate in the event (Table 4). The specific qualification deadlines for each event are provided in the *DTC Event Qualification Guide.*

Failing a previous qualification attempt does not preclude a team from resubmitting a revised qualification submission within the qualification deadlines for any given event. DARPA may adjust the qualification rules for each event and may choose to award qualification waivers for teams that have successfully participated in a prior Workshop or Challenge Event.

DARPA reserves the right to disqualify any team that is found to violate either the rules or applicable laws and regulations.

Event	Event Qualification	Event Date
Workshop 1	3/5/2024 - 4/5/2024	6/3/2024 - 6/8/2024
Challenge 1	6/28/2024 – 7/30/2024	9/28/2024 - 10/5/2024
Workshop 2	12/5/2024 -1/5/2025	3/17/2025
Challenge 2	5/28/2025 – 6/30/2025	**Data** 8/30/25 - Submission 10/4/2025 - Awards
Workshop 3	Winter 2025-2026	Winter 2025-2026
Challenge 3	Summer 2026	Fall 2026

Table 4 – Event qualification schedule.

7 DARPA Triage Challenge Technical Workshops

DARPA encourages vibrant information exchange and collaborative interactions among all DARPA Triage Challenge participants, to include DARPA technical staff, independent verification and validation (IV&V) teams, representatives from competitor teams, infrastructure developers, and other government partners. To that end, DARPA will host a workshop in each phase which will offer a forum for community building and cross-pollination of technical ideas and approaches as well as an opportunity for testing and integration.

In each phase (8 months into Phase 1, 4 months into Phases 2 and 3) DARPA will host a virtual workshop for the Data Competition, in which teams provide preliminary submissions for evaluation and receive performance results based on held-out data. The practice sessions will be followed by a 'lessons learned' discussion for all tracks and an opportunity to discuss real-world needs with Government partners.

As part of the workshops, teams will have the opportunity to confirm integration with the DARPA instrumentation and scoring systems to ensure compliant submissions ahead of the Challenge Events. Submissions for the workshops are not officially scored, but teams are encouraged to operate according to the Competition Rules to prepare for the Challenge events. Attendance at workshop events is required for all DARPA-funded teams. Self-funded teams may choose to attend.

We will hold a virtual lessons-learned meeting shortly after the workshop for teams to discuss experience gained regarding technical aspects of their systems at the workshop tests.

8 Human Subjects Research (HSR)

For the Data Competition, use of training data provided by DARPA does not constitute HSR, and competitors do not need to obtain IRB approval to use these data. However, DARPA-funded competitors require DARPA approval for the collection or use of any other human subject data. **Self-funded teams are prohibited from the collection or use of any other human subject data as part of their involvement in the DARPA Triage Challenge, beyond data and data-collection opportunities provided by DARPA, because DARPA HSR supervision is not feasible for teams not under DARPA contract.** Self-funded teams should carefully consider this limitation and should take this into account in their technical approach, leveraging other strategies as appropriate (*e.g.*, simulations).

DoD Definition of Human Subjects Research (HSR)

The term "human subject" can be applied to research efforts that meet EITHER of the following criteria: A living individual about whom an investigator (whether professional or student) conducting research:

- Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or
- Obtains, uses, studies, analyzes, or generates identifiable private information, personally identifiable information, or identifiable biospecimens.

Human Subjects Research involves:

- Activities that include both a systematic investigation designed to develop or contribute to generalizable knowledge and involve a living individual about whom an investigator conducting research obtains information or biospecimens through intervention or interaction with the individual, or identifiable private information, or biospecimens.

8.1 Handling of DARPA-provided data

Data Competition Datasets are provided by DARPA for use during DTC (henceforth dataset(s)). DARPA's mission requirement and intent are to safeguard privacy and civil liberties and the datasets have been intentionally de-identified to ensure—to the greatest extent practicable—that there is no reasonable basis to believe that the data could be used to trace a specific identity or present a risk of harm to any individual. Therefore, as previously acknowledged in the DTC Qualification process, competitors agree that they will *not intentionally* attempt to re-identify data in the datasets, nor attempt to download or share the datasets.

9 Secondary Triage: Data Competition Rules

9.1 Data - Illustrative Scenario

The objective of the Data Competition is to identify physiological signatures of injury derived from data captured by non-invasive sensors (contact-based or stand-off) to enable anticipatory decisions and prioritization for medical evacuation and care. Performers will develop algorithms that detect signatures in these data streams to provide decision support appropriate for austere and complex pre- hospital settings. Of particular interest are early signatures indicating need for LSIs against conditions that medics are trained and equipped to treat during secondary triage, such as hemorrhage and airway injuries.

The Data Competition is virtual only, and will use DARPA provided de-identified, multi-modal physiological data from trauma patients in diverse settings and cohorts provided by DARPA. DARPA will provide access to a subset of these data for algorithm training and evaluate competitor algorithms on held-out test data in workshops and end-of-phase challenge events. Challenge events will become progressively more complex and realistic from Phases 1 to 3.

9.2 Data - Technical Challenge Elements

The Challenge events will be designed to assess performance across various challenge elements, including: multiple data sources, multiple data inputs, raw data, and degraded data. The challenge elements are expected to become progressively more difficult from Phase 1 to Phase 3.

1. *Multiple data sources:* With varied source populations, patient demographics, types of injury,

and standard clinical operating procedures, each data set may have a different standard set of sensor readings, with different added sensors in each phase of DTC. Approaches must demonstrate robustness across a range of settings.

2. *Multiple data inputs*: Potentially including static data (e.g., mechanism of injury or anatomical injury pattern); multiple simultaneous, continuous streams of high-frequency waveforms; and point-of-care imaging data.

3. *Raw data*: Data as it comes from the sensors and health record systems (i.e., without any cleaning), with the noise, aberrant values, and dropouts that occur in clinical environments.

4. *Degraded data (year 2 and 3)*: DARPA also may inject additional challenges that can be expected in battlefield and civilian pre-hospital settings, such as severe degradation or total loss of a particular sensor, to test the robustness of competitor strategies to such plausible scenarios.

9.3 Data Types

Competitors in the Data Competition will be given data specifications for the training data that will be provided ahead of challenge events and used for evaluation. DARPA-provided datasets are comprised of data from two complementary academic hospital systems. The year 1 dataset contains around 3000 cases of pre-hospital and/or hospital data, the year 2 dataset is expected to contain around 2000 additional cases, and the dataset will continue to grow each year. Data sources include the following: discrete data (text or numeric values): patient characteristics, outcomes, procedures, labs; continuous data: (electrocardiogram) ECG, photoplethysmography (PPG), respiration rate (RR), heart rate (HR), SPO2, blood pressure (BP), Arterial-line BP. Additional data sources in year 2 include: ventilator data, pupillometry, and iSTAT portable blood analysis. Data sources are expected to increase with each phase of the challenge. Data dictionaries will be provided alongside each data release containing details on data format and proper interpretation.

9.3.1 Evaluation Data

While all data fields will be made available for training, only a subset of the data will be available at evaluation time for predicting LSIs. These will include data sources that would feasibly be available to medics *in situ* during treatment, excluding information about future events (e.g., outcomes) or contextual information typically collected post-treatment (e.g., medical history). In year 2, evaluation data may be further limited to encourage usage of physiological signals for LSI prediction. Details about data format and fields provided at evaluation time can be found in the Data Competition ICD.

9.3.2 Data Quality

DARPA-provided datasets contain deidentified data collected from trauma cases at two independent hospital systems. Due to the real-world challenges of medical data collection, there are varying degrees of consistency, completeness, and precision in data provided for this challenge. Minimal data cleaning has been applied to the data provided for training and used in evaluation, and teams are expected to develop their own mitigation strategies for handling real-world data.

9.4 Data - Scored Event Submissions

9.4.1 Solutions Submissions

For scored event submissions, teams must submit their solution in AWS ahead of the submission deadline. After the deadline, submissions will be built and evaluated using a held-out test dataset sampled from the

dataset. Teams will be provided with an opportunity to test their submissions ahead of the submission deadline to ensure their solutions are able to operate within the evaluation system. Guidelines on submission preparation and testing resources are included in the Data Competition ICD.

The solution submission window for the first challenge will open approximately 2 months prior to the Awards ceremony. Each qualified team must submit a single solution to be scored. The submissions will be evaluated and the final results will be announced alongside the Systems Competition results at the Competition Awards ceremony.

Challenge Event	Submission Window	Results Release
Challenge 1	*7/30/2024 – 8/30/2024*	*10/5/2024*
Challenge 2	*7/30/2025-8/30/2025*	*10/4/2025*
Challenge 3	*Summer 2026*	*Fall 2026*

Table 5 - Submission window for the Datal Competition teams.

9.4.2 Submission scoring

Each qualified team must submit a single solution per event to be scored. Team solutions will be evaluated under a simulated online prediction paradigm using held-out test data. Submitted solutions will be given incremental data within sequential time intervals that approximates streaming source data, and submission-predicted LSIs over time will be compared against ground truth LSIs appearing at later times in the case. The Event Score for each team is computed from performance metrics tailored to the prediction task. Details on scoring procedure and performance metrics are in Section 9.6 of this document.

Preliminary or partial results may be released to teams before the final ranking for review, however the final competition scores and any supporting materials will not be released until the event results are announced.

9.4.3 Human in the loop

The submitted solutions will be evaluated with no external operator interfaces such as command line inputs or user interventions. Teams are required to develop self-contained solutions that predict LSIs entirely autonomously without Human Supervisor interaction. Entries that require any user input or external commands will not be scored.

9.4.4 Run Termination

A scored run for a given case terminates upon any of the following conditions:

- **Time Expiration**: The processing time exceeded time limits, as described in Section 9.5.4.
- **Patient Expiration**: The patient expired or the case otherwise ended before another termination criterion was met.
- **Non-Compliant Submission:** The submission does not adhere to the ICD or otherwise fails at runtime.

9.4.5 Score Disputes

Score Disputes are intended to provide teams a mechanism to submit a formal dispute or request for review by the Chief Judge. All score disputes should be sent by email to the DARPA Triage Challenge email

address (triagechallenge@darpa.mil) within 24 hours of receiving data logs. All disputes or requests will be reviewed by the Chief Judge in a timely manner. All decisions made by the Chief Judge are final.

9.5 Data - Preliminary Event Scenarios

9.5.1 Preliminary Event Competition Environments

All computing operations and analysis involving the dataset must occur within the AWS ecosystem throughout the duration of the challenge. Each team will be provided an isolated, network-restricted AWS Workspace domain, which will allow teams to manage storage and computing resources. Team members will be provided Workspace user accounts and can instantiate computing resources (AWS EC2, SageMaker, etc.) on demand within these Workspaces (see Figure 3).

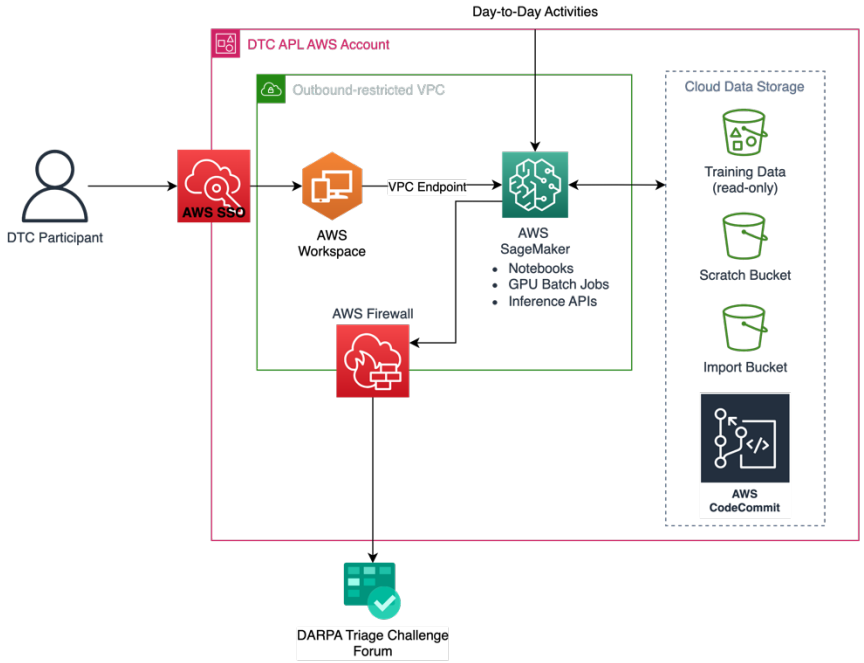

Figure 2 - DTC Participant AWS Architecture. Participants will perform all computations pertaining to sensitive data within the provided AWS ecosystem. Participants will be able to access the dataset stored in an S3 bucket, as well as instantiate computing resources (e.g., AWS EC2, AWS SageMaker, etc.) within their Workspace. Each team will be provided with a Scratch Bucket to share data or files among team members.

Teams would deploy AWS SageMaker instances to use GPU resources. Table 6 shows the available resources, their compute specifications, the maximum number of running instances per team, and the price per hour. The **ml.g4dn.4xlarge** instance type will be used for evaluation.

Instance Type	Compute Specifications				Purpose	Limit (per team)	Price per Hour ($)
	CPU	RAM (GB)	GPU	GPU Memory			
ml.t3.medium	2	4	0	0	General Purpose	15	0.05
ml.m5.large	2	8	0	0	General Purpose	15	0.115
ml.m5.2xlarge	8	32	0	0	General Purpose	15	0.461
ml.c5.xlarge	4	8	0	0	Compute Optimized	15	0.204
ml.c5.2xlarge	8	16	0	0	Compute Optimized	15	0.408

ml.c5.9xlarge	36	72	0	0	Compute Optimized	6	1.836
ml.r5.8xlarge	32	256	0	0	Memory Optimized	6	2.419
ml.r5.16xlarge	64	512	0	0	Memory Optimized	6	4.838
ml.p3.2xlarge	8	61	1	16	GPU - General	6	3.825
ml.p3.8xlarge	32	244	4	64	GPU - General	6	14.688
ml.g4dn.4xlarge	16	64	1	16	GPU - General	12	1.505
ml.g4dn.8xlarge	32	128	1	16	GPU - Training	6	2.72
ml.g5.4xlarge	16	64	1	24	GPU - Inference	1	2.03

*Table 6 - AWS processor options. **Bold** indicates instance type used for evaluation (ml.g4dn.4xlarge).*

Teams will be allocated a fixed budget at the beginning of each phase to spend on storage and computing resources. It will be the team's responsibility to manage their budget expenditures and resource consumption throughout the competition. All teams will be provided a dashboard to maintain and track their resources. Any remaining funds (up to 25%) at the end of a phase may be rolled into the subsequent phase. The Government's obligation for AWS funding under the DARPA Triage Challenge is subject to the availability of appropriated funds.

While working within Workspaces, participants are allowed to download data from the internet, allowing them to browse external websites and download open-source packages. However, participants are restricted from uploading data to ensure data does not leak outside the AWS ecosystem. Furthermore, if teams would like to use bespoke (privately developed) tools sitting outside AWS, they may package them and upload them to the Import Bucket, where they can retrieve them inside their SageMaker console.

9.5.2 Preliminary Event LSIs

As part of the dataset, treatments and related clinical actions will be identified and grouped into based on shared injury patterns and treatment paradigms. Each team in the Data competition is tasked with predicting the occurrence of these LSI groups, not the specific treatment or clinical action. For brevity, the LSI categories will be referred to as LSIs for the remainder of this document. LSIs with timestamps will be provided alongside the training and test datasets.

The list of LSIs for Phase 2 includes: airway & respiration, bleeding control, blood products, chest decompression, damage control procedures, and neurologic products & procedures. Details on the specific LSIs included in each group can be found in the data dictionaries provided with each data release. New datasets in Phase 2 will use the updated LSI list, and the Phase 1 dataset will be re-released for use in Phase 2 (see the DTC Community Forum for dataset release announcements).

9.5.3 Prediction Task

Submitted solutions will be evaluated within a simulated online prediction task, ingesting incremental continuous (e.g., physiological signals) and discrete (e.g., health record) data segments within sequential time windows. Given each data segment, the task is to predict the LSIs that the patient receives in the future relative to the observed data. Predictions are compared to the ground-truth LSIs occurring at future timepoints in the case. See Section 9.6 for details on the prediction task scoring criteria.

9.5.4 Time Limits

The evaluation duration is defined as the cumulative processing time for solutions to produce prediction responses across all test cases. To ensure timely evaluation and efficient solutions, DARPA will set an evaluation duration limit prior to each event as part of the Data Competition ICD. The time limit for Challenge 2 will be 48 hours, where the evaluation dataset contains 1273 cases with 13596 segments. The time limit applies to the sum of run times across all three runs in the phase 2 evaluation (see Section 9.6.7).

Additionally, a processing time limit will be enforced for each prediction equivalent to the duration of the data segment provided. For example, for an incremental data segment of 5 minutes duration, a prediction response is expected to be received within 5 minutes after the data is sent. If processing time to generate a prediction exceeds this time limit, the remainder of the case will be skipped in the evaluation and the case will be excluded from scoring.

9.6 Data - Scoring Criteria

Team will be evaluated based on accurate and timely prediction of future need for LSIs using physiological signals and contextual health information. Submitted solutions will be evaluated using a set of held-out test cases, where each case comprises pre- and in-hospital data from a single patient and hospital admission. Results for the Data Competition will be announced at the prize ceremony on the last day of the competition event.

9.6.1 Definitions

All solutions will be evaluated at a predetermined set of *evaluation timepoints* within each *case*, where a case encompasses the available pre- and in-hospital data and timestamped LSIs from a single hospital admission. At each evaluation timepoint, timestamped data within the *observation window* will be available to solutions to make predictions, where the observation window begins with the start of case and ends at the evaluation timepoint. The *minimum lead time* is the time before which an LSI needs to be predicted for actionable clinical predictions that would provide value for medical resource allocation and planning (see Section 9.6.4). The *prediction target* is then the set of ground-truth LSIs that occur within the *prediction window*, the time window beginning after the *minimum lead-time* following the evaluation timepoint (or end of the *observation window*) and extending to the end of case.

See Figure 3 for an illustration of the evaluation timepoint and surrounding windows, and how the prediction target is derived from the underlying LSI instances.

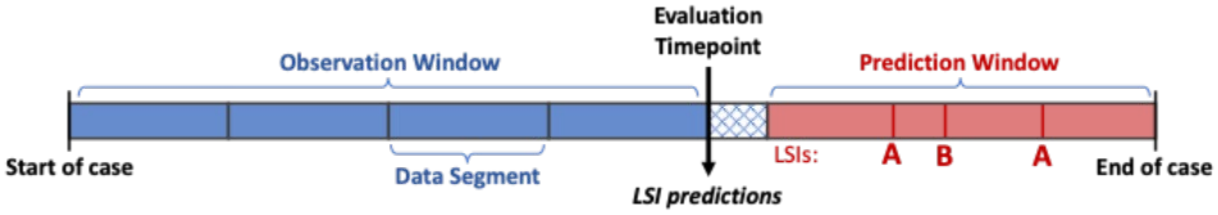

Figure 3 - Illustration of evaluation within a single case. Solutions are provided with data in sequential data segments containing timestamped health record and physiological signals data. The end of each data segment serves as an evaluation timepoint, where solutions generate predictions using data within the preceding observation window comprised of one or more data segments. LSI predictions are compared against the set of timestamped LSIs that occur in the future within the prediction window, excluding a predetermined minimum lead-time after the evaluation timepoint. In this example, the set of ground truth LSIs is [A, B], excluding any other LSIs that occur within the observation window or intervening gap (not pictured).

9.6.2 Online Prediction

To simulate online prediction, data within the observation window will be provided to solutions as a sequence of incremental data segments, with an evaluation timepoint at the end of each data segment.

Data segments contain timestamped health record and available physiological signals data within the segment start and stop timestamps. Note that all timestamps will be expressed as time elapsed relative to the beginning of the case. Data segments will be provided in sequential order within each case in its entirety with indicators for when one case ends and another begins.

In response to each data segment, team solutions will produce LSI predictions as a set of confidence scores to be compared against the ground truth LSIs occurring in the future relative to the observed data. Details on data segment and response formats are specified in the Data Competition ICD.

9.6.3 Prediction Time Horizon

To emphasize early prediction of medical need, team solutions will only be evaluated on predictions made within the first hour after hospital admission for LSIs received up to four hours after hospital admission. Using the terms defined above (and illustrated in Figure 3), the latest *evaluation timepoint* is 1 hour after hospital admissions, and the *prediction window* always ends at four hours after hospital admissions or at the end of care, whichever comes first.

Team predictions made after 1 hour following hospital admission will not factor into the evaluation, and predictions will only be evaluated against ground truth of LSIs received within four hours after hospital admission.

9.6.4 Minimum Lead-time

Similar to year 1, predictions must be made ahead of a minimum lead-time before the predicted event, representing the minimum time necessary for actionable clinical information that would provide value for medical resource allocation and planning. Predictions made after the minimum lead time preceding the ground truth LSI timestamp will be excluded from performance metrics and therefore will not contribute toward the team's Event Score. The minimum lead-time for year 2 will be 5 minutes.

9.6.5 Confidence Scores

Team solutions will provide LSI predictions in the form of confidence scores for each LSI group, which take values between 0 and 1. Confidence scores will be interpreted as the probability that the patient receives one or more treatments within the LSI group at a later time within the case. Confidence scores do not sum to 1 across LSI groups.

9.6.6 Performance Metric: Mean Squared Correct

Solutions in year 2 will be evaluated using a single metric, Mean Squared Correct (MSC), a modification of mean squared error that measures the similarity between LSI predictions and ground truth.

MSC is the weighted sum of squared magnitude of correct responses over time, where responses are confidence scores for LSI prediction at each time for each case. The weights are designed to address class imbalance in the dataset and incentivize equal error rates. Weights will be calculated from the held-back test dataset to be used in the next competition event and provided to teams ahead of the event within this document (see Appendix 1).

Let

- K be the set of LSI groups.
- $k \in \{1, 2, \ldots, K\}$ be an LSI group.
- I be the set of all cases in the evaluation set.
- $T_{i,k}$ be the number of evaluated segments for case i and LSI k, where $i \in I$ and $k \in K$. This excludes segments later than 1 hour after hospital admission and segments within the minimum lead time of an occurrence of LSI k.
- $t \in \{1, 2, \ldots, T_{i,k}\}$ be a segment for case i.
- Q be the set of data sources.
- $q_i \in Q$ be the source for case i.
- $y_{k,t,i} \in \{0,1\}$ be the ground truth presence or absence of LSI k in the future relative to segment t in case i.
- $\hat{y}_{k,t,i} \in [0,1]$ be the prediction response for LSI k at segment t in case i.

The unnormalized MSC for a single case i is:

$$MSC_i = \sum_{k=1}^{K} \sum_{t=1}^{T_{i,k}} w_{t,k,i} \left((1 - y_{t,k,i}) - \hat{y}_{t,k,i} \right)^2$$

The segment weight $w_{t,k,i}$ is defined as:

$$w_{t,k,i} = \frac{N}{n_{t,k,i}}$$

where:

- $N = \sum_{i \in I} \sum_{k \in K} T_{i,k}$ is the total number of evaluated responses across all cases and all LSIs.
- $n_{t,k,i} = \sum_{i' \in I} \sum_{k \in K} \sum_{t'=1}^{T_{i,k}} \mathbf{1}[y_{t',k,i'} = y_{t,k,i} \wedge q_{i'} = q_i]$ is the number of segments with ground truth value $y_{t,k,i}$ for LSI k from the same source q_i.

The raw unweighted MSC (i.e., the squared difference term in the equation above) takes a value of 1 if the prediction matches ground truth, and 0 if the prediction is the opposite of ground truth. See Figure 4 for an illustration of the unweighted MSC calculation.

Figure 4 - Illustration of confidence score prediction and unweighted Mean Squared Correct (MSC) metric. Predictions ($\hat{y}_t$, closed circles) and ground truth (y_t, open circles) over time are shown simplified for a single case and single LSI (subscripts omitted). Before the LSI occurrence (dotted line), ground truth is 1, whereas after LSI occurrence, ground truth is 0. MSC is the weighted sum of squared magnitudes of red lines at each prediction. Note that the exclusion of the prediction within the minimum lead time is not depicted in this figure (see Section 9.6.5).

The segment weight $w_{t,k,i}$ scales the raw unweighted MSC according to how common it is in the test dataset: the segment weight is higher for less common segments, and lower for more common segments. The segment weight is calculated as the inverse frequency of segments within the test dataset with the

same ground truth $y_{t,k,i}$ for the same LSI k from the same data source as case i.

$$w_{t,k,i} = \frac{Total\ count\ of\ segments}{Count\ of\ segments\ with\ same\ ground\ truth\ y_{t,k,i}\ for\ LSI\ k\ from\ same\ source\ as\ case\ i}$$

Counts only include segments for which predictions are evaluated, namely segments up to 1 hour after hospital admission and outside of the minimum lead time of an LSI occurrence. Weights for Challenge Event 2 can be found in Appendix 1, along with additional details on their calculation. For subsequent competition events, weights will be updated in future versions of this document.

Note that the unnormalized case MSC will have a different scale depending on the case. To compare MSC between cases, use the normalized case MSC:

$$\widehat{MSC}_i = \frac{MSC_i}{\sum_k \sum_t w_{t,k,i}}$$

While only the unnormalized case MSC is used directly in the Event Score calculation (with normalization occurring later, as described below), the normalized case MSC may be useful for other analyses involving case-to-case comparisons.

9.6.7 Multiple Runs and Run Score

To offer teams the opportunity to demonstrate algorithm performance under more austere settings, team solutions will be evaluated in multiple runs using the same set of held-out test data with further limited data sources available at test time. Data has been divided into the following categories:

- Basic EHR: demographics, injury, GCS, EHR vitals, instantaneous labs (i.e., iSTAT)
- Expanded EHR: injury details, labs, procedures, medications, additional context
- LSI: LSI information needed for prediction.
- Vitals: Continuous waveforms and trends of vital signs

The following run schedule lists predictor data available at test time for each run by category:

Run 1. Basic EHR, Expanded EHR, LSI, Vitals.
Run 2. Basic EHR, LSI, Vitals.
Run 3. LSI, Vitals.

More details about data fields provided during each run can be found in the Data Competition ICD.

The Run Score MSC_r for run r is the normalized sum of MSC over cases:

$$MSC_r = \frac{\sum_i MSC_{i,r}}{\sum_i \sum_k \sum_t w_{t,k,i}}$$

where $MSC_{i,r}$ is the case MSC for case i and run r (see equation in previous section, where r was omitted for brevity), and $w_{t,k,i}$ is the segment weight described above. After normalization, the Run Score takes values between 0 and 1.

9.6.8 Event Score

The Event Score is the weighted sum of Run Scores across runs:

$$Event\ Score = \sum_{r} w_r\, MSC_r$$

where MSC_r is the Run Score for run r, weighted by the run weight w_r. The run weight will sum to 1 across runs, so the Event Score takes values between 0 and 1. For Challenge Event 2, the three runs will be equally weighted, $w_r = 1/3$.

9.6.9 Final Ranking

The final ranking will be determined by decreasing Event Score. If multiple teams have an identical Event Score, the team with a higher MSC summed across cases in the most austere run (e.g., Run 3 above) will be ranked higher.

9.6.10 Minimum Benchmarks to Win Prizes

In phase 2 there is a minimum benchmark for winning prizes. Self-funded teams must achieve each of the below performance requirements in at least one evaluation run:

- Exceed the Event Score of baseline algorithms provided by the DTC IV&V team.
- Exceed 0.65 in average of sensitivity and specificity across segments for the following LSI categories 1) Airway and Respiration and 2) Bleeding Control. The threshold for calculating specificity and sensitivity will be selected such that it maximizes the average of sensitivity and specificity.

10 Appendix 1: Segment Weights

Segment weights were calculated from the held-out dataset to be used in Challenge Event 2, including held-back cases from both the Phase 1 and the Phase 2 datasets.

To implement the formula above and calculate segment weights, the following procedure was used:

1. Start with inventory of data segments, including data source and ground truth for each LSI.
2. Exclude segments with end time later than 1 hour after hospital admission.
3. Group segments by data source, LSI, and ground truth value (0 or 1).
4. Within each group with ground truth value 1, if duration from segment end time to last LSI timestamp within the case is less than minimum lead time, exclude segment from group.
5. Let group frequency be the size of each group divided by the total number of segments.
6. Define segment weight as inverse of the group frequency.

Tables 7 and 8 contain the segment weights to be used in Challenge Event 2. These weights will be provided in metrics software in AWS as part of the *client-shell*, along with code used to calculate them (see DTC Forum for announcements about software releases).

LSI group k	$y_{t,k,i}$	$w_{t,k,i}$
Airway & Respiration	0	25.27
	1	342.32
Bleeding Control	0	23.59
	1	2339.71
Blood Products	0	25.65
	1	275.73
Chest Decompression	0	24.07
	1	816.92
Damage Control Procedures	0	24.26
	1	619.51
Neurologic Products & Procedures	0	23.93
	1	1008.33

Table 7. Segment Weights for cases from the UMB dataset.

LSI group k	$y_{t,k,i}$	$w_{t,k,i}$
Airway & Respiration	0	74.84
	1	303.13
Bleeding Control	0	60.28
	1	4155.0
Blood Products	0	64.92
	1	753.09
Chest Decompression	0	63.17
	1	1012.56
Damage Control Procedures	0	63.43
	1	930.46
Neurologic Products & Procedures	0	59.80
	1	8310.0

Table 8 – Segment Weights for cases from the Pitt dataset.

11 DTC Glossary

Chief Official – Program manager or higher DARPA authority for the DARPA Triage Challenge.

Systems Competition –Primary Triage Competition run with actors on a real course (Track A, B).

Data Competition – Secondary Triage Competition (Track D, E).

Chief Judge – DARPA-designated individual with the sole and final authority to make any decisions related to the rules or scoring.

Judge – DARPA-designated individual with authority to make decisions related to rules, scoring, and safety, with decision-making authority only superseded by the Chief Judge.