FOUNDATION MODEL OF ELECTRONIC MEDICAL RECORDS FOR ADAPTIVE RISK ESTIMATION

PREPRINT (UNDER REVIEW)

Pawel Renc[1,2,3], Michal K. Grzeszczyk[1,2], Nassim Oufattole[4], Deirdre Goode[5,2], Yugang Jia[4], Szymon Bieganski[6], Matthew B. A. McDermott[2], Jaroslaw Was[3], Anthony E. Samir[7,2], Jonathan W. Cunningham[8,2], David W. Bates[8,9,2], and Arkadiusz Sitek[1,2,*]

[1]CAMCA, Massachusetts General Hospital, Boston, USA
[2]Harvard Medical School, Boston, USA
[3]AGH University of Krakow, Krakow, Poland
[4]Massachusetts Institute of Technology, Cambridge, USA
[5]Newton Wellesley Hospital, Newton, USA
[6]Medical University of Lodz, Lodz, Poland
[7]CURT, Massachusetts General Hospital, Boston, USA
[8]Brigham and Women's Hospital, Boston, USA
[9]Harvard Chan School of Public Health, Boston, USA
[*]*Corresponding author: sarkadiu@gmail.com*

March 17, 2025

ABSTRACT

Background, The U.S. allocates nearly 18% of its GDP to healthcare but experiences lower life expectancy and higher preventable death rates compared to other high-income nations. Hospitals struggle to predict critical outcomes such as mortality, ICU admission, and prolonged hospital stays. Traditional early warning systems, like NEWS and MEWS, rely on static variables and fixed thresholds, limiting their adaptability, accuracy, and personalization; **Methods,** We developed the Enhanced Transformer for Health Outcome Simulation (ETHOS), an AI model that tokenizes patient health timelines (PHTs) from EHRs and uses transformer-based architectures to predict future PHTs. The Adaptive Risk Estimation System (ARES) leverages ETHOS to compute dynamic, personalized risk probabilities for clinician-defined critical events. ARES also features a personalized explainability module that highlights key clinical factors influencing risk estimates. We evaluated ARES on the MIMIC-IV v2.2 dataset in emergency department settings, benchmarking its performance against traditional early warning systems and machine learning models; **Results,** From 299,721 unique patients, 285,622 PHTs (60% with hospital admissions) were processed, comprising over 357 million tokens. ETHOS outperformed benchmark models in predicting hospital admissions, ICU admissions, and prolonged stays, achieving superior AUC scores. Its risk estimates were robust across demographic subgroups, with calibration curves confirming model reliability. The explainability module provided valuable insights into patient-specific risk factors; **Conclusions,** ARES, powered by ETHOS, advances predictive healthcare AI by delivering dynamic, real-time, personalized risk estimation with patient-specific explainability. Its adaptability and accuracy offer a transformative tool for clinical decision-making, potentially improving patient outcomes and resource allocation. We release the full code at `github.com/ipolharvard/ethos-ares` to facilitate future research.

1 Introduction

The United States allocates nearly 18 percent of its GDP to healthcare, yet Americans have shorter lifespans and poorer health than residents of other high-income nations. Among these countries, the U.S. not only has the lowest life expectancy but also the highest rates of preventable deaths [1]. Hospitals face mounting challenges managing patient influx and identifying individuals at risk for critical outcomes, including mortality, intensive care unit (ICU) admission, or prolonged hospital stays [2]. Accurate prediction of critical clinical events is essential for enhancing patient care and optimizing the timely allocation of limited healthcare resources [3]. Early identification of at-risk patients enables clinicians to prioritize interventions, anticipate potential escalations in care, and improve outcomes while simultaneously reducing costs [4, 5]. However, current methodologies often fail to fully utilize the vast and complex data available in electronic health records (EHRs), a limitation that becomes particularly evident in emergency settings where time-sensitive decisions are critical [6, 7, 8, 9, 10]. Traditional scoring systems, such as the National Early Warning Score (NEWS) [11] and the Modified Early Warning Score (MEWS) [12], rely on static variables and predefined thresholds, constraining their ability to adapt to dynamic and multifaceted patient data. Similarly, conventional machine learning models depend on preselected predictors of patient deterioration, requiring the inclusion of only a limited number of variables. These approaches are further hindered by their reliance on specific cutoff points for data inclusion (e.g., triage, 24-hour windows), which can overlook valuable longitudinal patterns.

Recent advances in generative machine learning, particularly in transformer architectures [13, 14, 15, 16], which underpin the success of Large Language Models (LLMs) [17, 18], have unlocked unprecedented capabilities for processing high-dimensional, heterogeneous, time-stamped, and episodic health data derived from electronic health records (EHRs) [14, 19, 20, 21, 22, 23]. In this work, we build on our previous development, the Enhanced Transformer for Health Outcome Simulation (ETHOS) [14]. While our approach shares some similarities with the works of [20] or [19] it differs in how EHR data are encoded and processed by the transformer model. ETHOS is designed to provide zero-shot predictions; once the model is trained, no additional fine-tuning is required to make inferences. The model operates on Patient Health Timelines (PHTs), which are tokenized sequences of events extracted from EHRs, including demographics, medical history, medications, and more (the full list is provided in Table S7). Using known PHTs up to a given point in time, ETHOS predicts future PHTs (Figure 1). ETHOS enables dynamic, real-time risk assessment by computing a range of possible Patient Health Timelines (PHTs) for a defined outcome, such as ICU admission. If the probability of an adverse event exceeds a critical threshold, appropriate interventions can be initiated to mitigate risk. This continuous probability estimation functions as an early warning system, similar to an experienced physician's intuition in identifying high-risk patients. Unlike traditional models that require predefined inference tasks, ETHOS operates as a single, unified model, allowing for the simultaneous assessment of multiple positive and negative outcomes without retraining. Probabilities for various clinical events are dynamically updated as new patient data becomes available, ensuring adaptability throughout the care process. We refer to this flexible and scalable risk prediction framework as the Adaptive Risk Estimation System (ARES), as illustrated in Figure 2.

In this paper, we present the development of ARES and introduce a novel explainability framework that delivers fully personalized insights, allowing clinicians to understand the specific factors influencing the system's risk predictions for individual patients. We benchmark the performance of ARES against state-of-the-art methods across multiple clinically relevant tasks, demonstrating its superior predictive accuracy. Using Emergency Department (ED) datasets from MIMIC-IV-ED [24, 25, 26], we validate its effectiveness and provide the accompanying code for the full reproduction of all the experiments by other researchers.

2 Methods

2.1 Data

In this study, we used the Medical Information Mart for Intensive Care (MIMIC-IV) version 2.2 database [24, 25], including its ED extension. MIMIC-IV, developed by the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center contains de-identified health records for almost 300,000 patients either admitted to the ED and/or hospital at BIDMC from 2008 to 2019. Detailed patient demographics are presented in Table S2. We extracted relevant data from the MIMIC-IV tables as detailed in Table S7. Laboratory tests and medications were standardized using ATC codes, and all diagnostic and procedural codes were mapped to ICD-10 when necessary, as described in detail [14]. Additional tables requiring advanced processing, such as clinical notes, were not included in the current implementation of ETHOS.

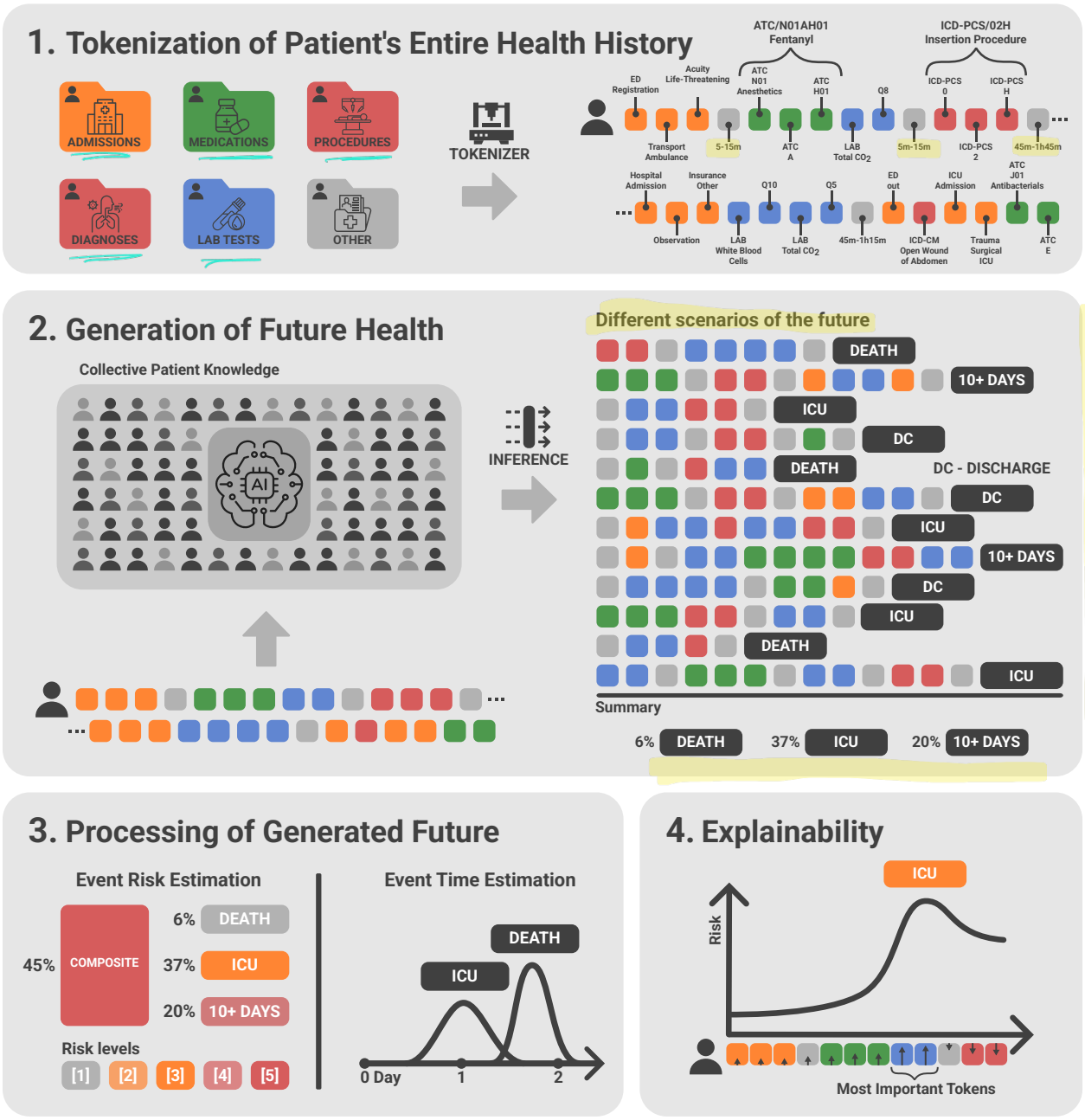


Figure 1: **Workflow of the Adaptable Risk Estimation Score (ARES) Framework.** This figure illustrates the ARES framework, developed on the ETHOS model, for dynamic and explainable risk evaluation. Panel 1 depicts the tokenization of a patient's entire health history into structured events represented as a sequence of tokens (PHTs), incorporating standardized coding systems such as ATC for medications, ICD-PCS for procedures, and others. Panel 2 demonstrates how the ETHOS model trained on a large dataset of PHTs to simulate potential future patient health timelines (fPHTs). By analyzing a particular patient's known PHT and generating multiple fPHTs, the model estimates the probabilities of critical outcomes, such as inpatient death, ICU admission, or a prolonged hospital stay exceeding 10 days. Panel 3 showcases the result of processing of fPHTs to calculate event-specific risks and predict the timing of these events, should they occur. Risk levels are defined across five categories, color-coded for enhanced clinical interpretability. Panel 4 showcases the explainability module, which identifies the key factors influencing specific risk estimates, offering personalized and actionable insights to support clinical decision-making. In this example, blue tokens indicate factors contributing to an increased risk of ICU admission.

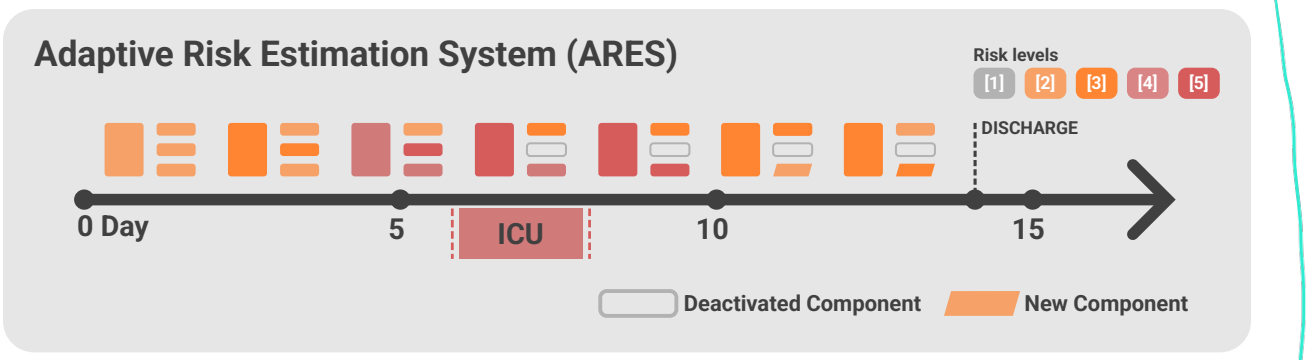


Figure 2: **Timeline of a Patient's Hospital Stay and Hypothetical Risk Predictions by ARES.** This figure illustrates the timeline of a patient's hospital stay, from admission to discharge around Day 14, demonstrating how ARES dynamically adjusts its predictions based on the patient's evolving clinical status and medical history. By Day 5, ARES predicts a high risk of ICU admission, which is subsequently confirmed as the patient is admitted around Day 6. Once the patient is in the ICU, ARES discontinues ICU risk evaluation, as indicated by the "Deactivated Component" label. After the ICU stay, ARES identifies an increased likelihood of a hospital stay exceeding 10 days. Upon reaching the 10-day threshold, ARES automatically recalibrates its predictions, replacing the previous risk estimation with the likelihood of a 15-day stay, now categorized as a "New Component" in the risk assessment.

2.2 Tokenization, PHT Construction, model training

The core of ETHOS lies in constructing PHTs from electronic medical records (EMRs) using a tokenization strategy that captures diverse clinical events. A PHT represents a patient's medical history as a sequence of tokens, each encoding specific health-related information organized chronologically. This structured representation enables comprehensive modeling of patient journeys and more accurate clinical predictions. To build PHTs we used the MEDS-DEV [27] extraction pipeline that converts EHR data to an intermediate format called MEDS [28] to facilitate further data transformations. Advanced transformation operations were subsequently applied, breaking down each event into 1 to 7 tokens based on its complexity. Simpler events required fewer tokens, while intricate ones, such as multi-component lab results, were represented with more tokens to encapsulate their detailed information.

For example, lab test results were encoded using quantile-based tokens to represent clinical significance. Time-interval tokens were added to mark the elapsed time between successive events, with intervals shorter than 5 minutes omitted and longer gaps tokenized into 19 distinct interval tokens. Continuous numerical values, such as lab test results, were similarly quantile-encoded using ten quantiles, balancing clinical interpretability and predictive precision. Diagnostic and procedural codes, including ICD-10-CM, ICD-10-PCS, and ATC drug codes, were encoded hierarchically, which leveraged their inherent structure to enhance the transformer model's attention mechanisms. For more details, refer to [14].

Static patient attributes such as gender, marital status, race, and body mass index (BMI) were encoded using a single token depending on the value. For age, tokens of quantiles were reused, allowing age representation from 0 to 99. For instance, a 46-year-old patient would be coded as Q5 and Q7. Attributes with potential variability were represented using their most recently known value at the start of the timeline. By incorporating these elements, ETHOS ensured a richer and more adaptable representation of patient timelines.

The dataset was split into two disjoint groups: training/validation (90%) and testing (10%). During the training phase, 6 million tokens (1.8% of the entire dataset) were used for validation to balance model optimization and computational efficiency. The detailed statistics about the tokenized dataset are available in Table S6 and S9, and information about the model is in S1.

2.3 Probabilistic inference

The ETHOS model generates probabilities of future clinical events by leveraging tokenized PHTs and employing a transformer-based generative model. During inference, ETHOS autoregressively generates tokens, each representing a potential future event, until predefined stopping conditions are met, such as the appearance of a token of interest or meeting the simulation time limit. By simulating multiple future PHTs (fPHTs) for each patient, ETHOS accounts for inherent uncertainties and produces robust probabilistic predictions. For example, if N simulated fPHTs are generated and M indicates inpatient mortality, the estimated probability of mortality is calculated as M/N (Appendix A). This

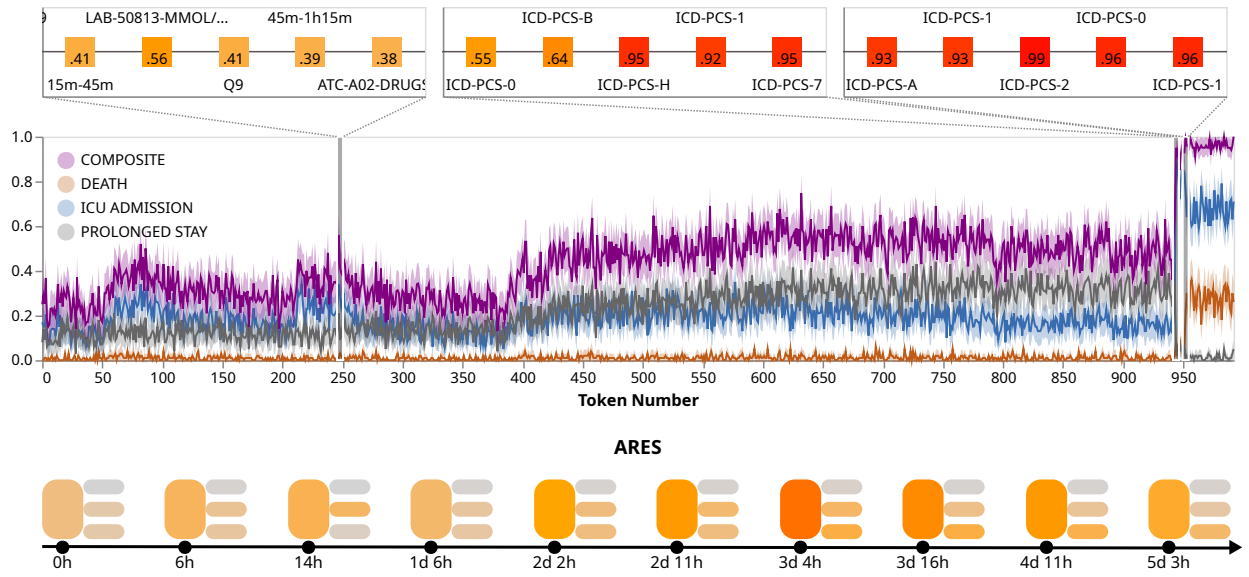


Figure 3: **ARES Risk Trajectories.** This figure illustrates risk trajectories for nearly 1000 tokens preceding patient death, as monitored by ARES, which evaluates the probability of death, ICU admission, prolonged hospital stay, and a composite risk score. The lower panel provides a color-coded representation of risk with the actual time since the ED presentation. In contrast, the upper panel highlights three 5-token regions influencing risk predictions at areas marked by the thin gray bar. In the first region, token LAB-50813 (Lactate Blood Test) increases the composite risk score from 0.41 to 0.56, but since the result falls in Q9 (80–90th percentile), ETHOS downgrades the risk estimate back to the previous level. In the second region (close to the end), a sharp increase in composite risk occurs due to heightened ICU admission triggered by ICD-PCS code 0BH17EZ, which is coded by 7 tokens (only 5 visible), which represents Endotracheal Airway Insertion into the Trachea via Natural or Artificial Opening. The 'H' token specifically signals ETHOS to escalate the ICU risk to nearly 1.0, indicating that the patient is being intubated de novo. The ICD-10-PCS breakdown confirms the procedure as a respiratory intervention involving tracheal insertion via a natural or artificial opening. ICD-PCS 0BH17EZ does not increase the risk of death, but the next ICD-PCS 5A12012 (5 tokens coding A1202 visible) raises the risk of death to about 0.25. We note that an increased risk of death is associated with a decreased risk of ICU admission, as these are competing risks. This visualization demonstrates how ARES dynamically adjusts risk scores based on evolving patient data, integrating clinical trajectories into real-time risk assessment.

stochastic, scenario-based methodology enables comprehensive modeling of patient trajectories, facilitating precise and dynamic risk assessments tailored to individual patient profiles.

2.4 Explainability

As illustrated in Figure 3, stochastic simulations can be initiated not only from the most recent token representing current information but also from any preceding token in the patient's history. This allows risk estimates to be visualized as a time series, highlighting how specific medical events affect risk over time. This approach provides intuitive visualizations, offering clinicians clear insights into the factors contributing to current risk values.

2.5 Methods used for benchmarking

We followed benchmarking tasks for emergency department models presented in the Emergency Department MIMIC-IV-ED benchmark paper [29]. Three tasks were defined: prediction of the hospital admission at triage, prediction of the critical outcome (death or transfer to ICU within 12 hours) at triage, and ED re-presentation within 72 hours after discharge from ED. We applied machine learning methods (logistic regression, random forest, gradient boosting), scoring systems MEWS [12], NEWS [11, 30, 31], Rapid Emergency Medicine Scores (REMS) [32], Cardiac Arrest Risk Triage (CART) [33], five-level triage system Emergency Severity Index (ESI) [34] and neural networks-based models including multilayer perceptron, Med2Vec [35] and Long Short-Term Memory (LSTM) [36].

To compare tasks used for early warning scores, we compared the MEDS-Tab library [37] which was used to establish a baseline. MEDS-Tab converts time-series EHR data into a tabular format by aggregating features across multiple time

windows. It takes longitudinal patient data and applies various aggregation functions (like sum, count, min, max) over different historical window sizes to create fixed-size feature vectors, where each feature represents a combination of a medical code, time window, and aggregation method. XGBoost [38] models are trained on these tabular features computed from data windows prior to each prediction time point for each clinical task.

2.6 Statistical Methods

The performance of predictive models was evaluated using Receiver Operating Characteristic (ROC) curves and corresponding Area Under the Curve (AUC) values. Bootstrapping techniques were employed to estimate 95% confidence intervals (CIs) for AUCs. Model predicted probabilities were compared with observed event frequencies using calibration curves to evaluate ETHOS's reliability and alignment with real-world clinical outcomes. All statistical analyses were conducted using Python-based libraries, including scipy and scikit-learn [39, 40]. Data visualization, including ROC curves, calibration plots, and other statistical figures, was performed using matplotlib, seaborn and altair.

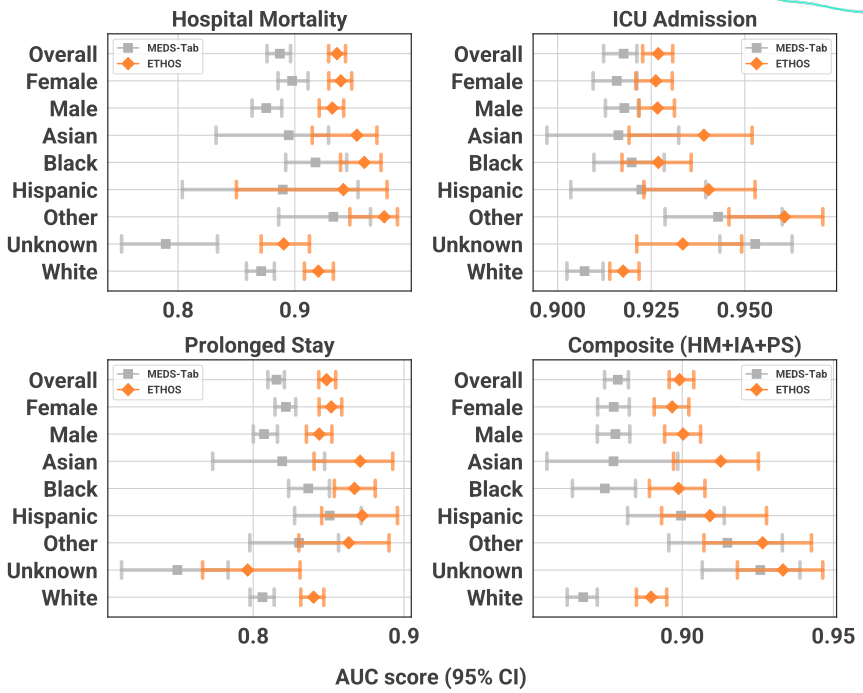


Figure 4: **AUC Comparison Between ETHOS and MEDS-Tab Across Demographic Subgroups and Prediction Tasks.** AUC scores with 95% confidence intervals are shown for ETHOS (orange) and MEDS-Tab (gray) across four prediction tasks: Hospital Mortality, ICU Admission, Prolonged Stay, and Composite Outcome (Hospital Mortality + ICU Admission + Prolonged Stay). Performance is reported for the overall population and stratified by gender (Female, Male) and race (Asian, Black, Hispanic, Other, Unknown, White). ETHOS consistently outperforms MEDS-Tab across all demographic subgroups and tasks.

3 Results

Following the tokenization process, the data of 299,721 unique patients from the MIMIC-IV dataset was converted into 285,622 PHTs, which were subsequently used for training and testing. The discrepancy arises from some patients lacking associated data after tokenization. Of the total PHTs, approximately 60% (180,733) contained hospital admissions records. The tokenized dataset comprised over 357 million tokens in total. Detailed information regarding the MIMIC-IV data used, patient demographics, characteristics of the PHTs and tokens, and descriptive statistics are provided in supplementary data (Table S2,S7,S6,S9). The model was trained and validated on 90% of the PHTs, with the remaining 10% reserved for testing. During inference, at least 100 fPHTs were generated for each investigated task.

The predictive performance of ETHOS and MEDS-Tab was evaluated for four critical clinical outcomes: hospital mortality, ICU admission, prolonged hospital stay, and a composite risk score (HM+IA+PS). Prolonged stay was defined as a stay longer than 90th percentile of all stays. All predictions were performed at patient admission. As

summarized in Figure 4, 5, and Table S1, ETHOS consistently outperformed MEDS-Tab across all outcome measures, demonstrating superior AUC values. Notably, ETHOS yielded higher AUC values across all racial groups, with the most significant improvement observed among Asian and Hispanic patients. The model's robustness across diverse populations suggests its potential for mitigating disparities in predictive accuracy.

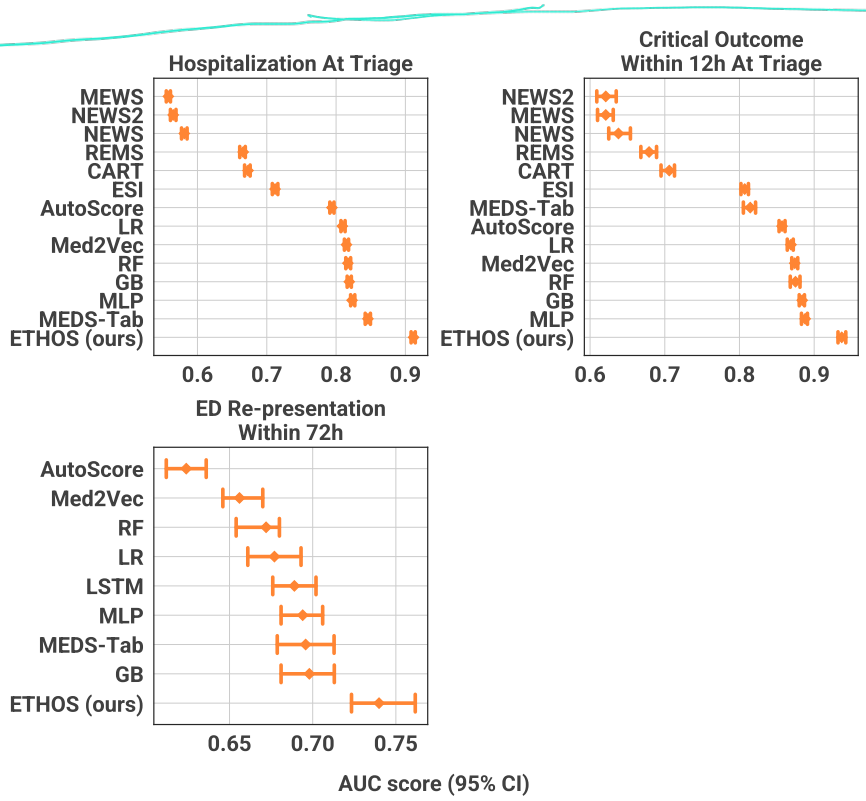


Figure 5: **Predictive results for the ED benchmark tasks.** Fewer methods appear in the ED re-presentation task (right) because score-based approaches, designed specifically to estimate in-hospital deterioration, are not applicable once the patient has left the ED. Ethos consistently achieves the best performance across all evaluated tasks.

Figure 3 illustrates the dynamic risk trajectories generated by ARES, showcasing how the system continuously updates probability estimates for key clinical outcomes, including ICU admission, prolonged hospital stay, and mortality, as new clinical events occur. The figure highlights specific medical interventions, such as laboratory tests and procedures, that drive significant changes in risk estimates, demonstrating ARES's ability to integrate evolving patient data into real-time risk assessment. The results underscore the model's capacity to capture complex temporal relationships between clinical events, dynamically recalibrating risk scores based on patient status and treatment progression.

In addition to risk which are part of ARES and to contextualize the predictive capabilities of ETHOS, we compared its performance against traditional early warning scores and other ML models. Figure 5 presents the AUC values (ROC curves in Figure S2) for key ED benchmark tasks: hospitalization at triage, critical outcomes within 12 hours of triage, and ED re-presentation within 72 hours post-discharge. ETHOS demonstrated consistently superior predictive accuracy across all evaluated tasks. We provide detailed numerical values in supplementary data (Table S3,S4,S5).

The risks provided by ETHOS were also found to be well-calibrated, as tested by calibration curves. Brier scores were found in the range 0.01-0.14 depending on the task, indicating excellent to good performances, as shown in Figure S3.

4 Discussion

The ARES framework introduces an innovative approach to building predictive models by leveraging cutting-edge artificial intelligence technology. Several aspects of this approach distinguish it from traditional models. First, ARES enables dynamic risk estimation at any time during a patient's stay, from admission to discharge. Powered by ETHOS [14], ARES utilizes PHTs and incorporates all available clinical information at the time of risk estimation. Unlike traditional models, which rely on static data points such as information collected within 24 hours after admission

or ED presentation or data up to triage [29, 41], ARES continuously adapts to the patient's evolving clinical status. This adaptability overcomes a key limitation of static models, which may not perform optimally outside the narrow time frames for which they are designed. This capability is demonstrated in the accompanying Figure 3 and Table S8, which illustrate how risk evolves over time during a patient's hospital stay. These visualizations, which depict how personalized risk evolves over time to reach the current estimates, provide insights into the specific factors driving model predictions for each patient. They highlight clinical events associated with increased or decreased risk, offering real-time explainability. By identifying the most influential features contributing to an individual's risk assessment, ARES empowers clinicians with a clearer understanding of the rationale behind each prediction.

As illustrated in Figure 1, ARES can estimate risk for various critical events, such as in-hospital mortality, ICU admission, and prolonged hospital stays. Beyond these standard metrics, additional indicators can be integrated seamlessly, including the risk of ICU admission during a specific length of stay, ICU readmission, acute kidney injury, sepsis, cardiac arrest, or 30-day readmission, and others. The ETHOS model, which underpins ARES, allows for the dynamic combination of these risks into composite measures while accounting for their interdependencies. For example, the occurrence of mortality on Day 8 would render the probability of a 10-day hospital stay zero. This ability to incorporate conditional and causal relationships between tracked events is another strength of ARES. Importantly, integrating additional metrics does not require model retraining or modifications to the ETHOS model. Once a range of possible future PHTs has been generated, any additional metrics can be calculated with minimal computational resources, making ARES scalable and adaptable to diverse healthcare settings.

In its current implementation, ETHOS distills multiple fPHTs into a single predictive decision, such as inpatient mortality. However, this approach overlooks the wealth of longitudinal information contained in these trajectories, including the sequence of clinical events that lead to a particular outcome, or the absence thereof. By merely predicting the likelihood of an adverse event, valuable insights into the pathways that contribute to deterioration or recovery remain underutilized. Expanding ETHOS to provide a more granular, trajectory-based interpretation of risk would allow clinicians not only to assess a patient's probability of experiencing a critical event but also to understand the evolving clinical course leading to that outcome including the cost. This enhanced approach would address a key limitation highlighted in the early warning paradox [42], where models trained on retrospective data may fail to capture the full complexity of clinical interventions and their effects on patient outcomes. Moving forward, we aim to refine ETHOS to incorporate and visualize these probabilistic trajectories. This will equip clinicians with deeper, more actionable insights into clinical risk dynamics and potentially provide new information about causality in patient outcomes.

This study has limitations. ETHOS was demonstrated using PHTs derived from MIMIC-IV-ED data, and its direct applicability to data from other institutions may be limited without retraining using additional data from other institutions. Electronic medical record (EMR) systems vary significantly across institutions, influenced by differences in clinical practice, care pathways, patient populations, and geographic location. These variations can impede the direct transferability of AI models trained on one dataset to another. In certain applications, such as radiology or pathology, data inputs like medical images are relatively standardized, allowing models trained in one institution to perform well in others. However, EMR data pose unique challenges due to their variability. Models trained on data from one institution may produce inaccurate risk estimates when applied to data from another, particularly if clinical practices differ [43]. To mitigate this limitation, the model code for ETHOS-ARES is compatible with the MEDS [28] health AI data standard, making it easier for other researchers to reliably train identical model architectures on their local data.

Data standardization is often proposed as a solution to address the challenges of variability in healthcare data. However, achieving meaningful standardization would require identifying commonalities between healthcare systems, an endeavor that may not be feasible given the diversity of clinical practices, patient populations, and institutional workflows. In our view, a more robust approach is to train AI models, such as ETHOS, on raw data from diverse institutions, allowing the model itself to learn and interpret the underlying patterns and clinical pathways. This approach mirrors the capability of large language models (LLMs) to discern meaning from vastly different styles of text and presentations or even different languages, leveraging the same foundational transformer architecture as ETHOS.

In summary, artificial intelligence advancements have unlocked unprecedented opportunities for innovative solutions like ARES, which leverage large amounts of heterogeneous data to develop general-purpose models with superior predictive power compared to state-of-the-art methods. ARES not only enables dynamic, personalized risk estimation but also provides real-time explainability, empowering clinicians to make more informed decisions. Furthermore, its modular design and the underlying ETHOS model allow for seamless integration of additional data types, such as radiology, genetics, and other institutional datasets, paving the way for even greater predictive accuracy and applicability in diverse healthcare settings.

As healthcare costs and complexity continue to rise, PHT-based frameworks like ARES show a promising pathway towards data-driven AI-enabled individualized patient care with the potential to reduce morbidity, improve outcomes, and lower healthcare costs.

Acknowledgments

We thank Kinga Renc, M.Arch, for her invaluable assistance with graphic design. This work was supported in part by National Institutes of Health (NIH) grant number HL159183.

References

[1] Munira Z Gunja, Evan D Gumas, and Reginald D Williams II. *U.S. Health Care from a Global Perspective, 2022: Accelerating Spending, Worsening Outcomes*. Jan. 2023. (Visited on 01/26/2025).

[2] Committee on the Future of Emergency Care in the United States Health System. *Hospital-based emergency care: at the breaking point*. Washington, D.C., DC: National Academies Press, 2007.

[3] Kum Khiong Yang et al. "Managing emergency department crowding through improved triaging and resource allocation". en. In: *Oper. Res. Health Care* 10 (Sept. 2016), pp. 13–22. DOI: 10.1016/j.orhc.2016.05.001.

[4] Devin J Horton et al. "Modified early warning score-based clinical decision support: cost impact and clinical outcomes in sepsis". en. In: *JAMIA Open* 3 (2 July 2020), pp. 261–268. DOI: 10.1093/jamiaopen/ooaa014.

[5] Roy Adams et al. "Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis". en. In: *Nat. Med.* 28 (7 July 2022), pp. 1455–1460. DOI: 10.1038/s41591-022-01894-0.

[6] Dana P Edelson et al. "Early warning scores with and without artificial intelligence". en. In: *JAMA Netw. Open* 7 (10 Oct. 1, 2024), e2438986. DOI: 10.1001/jamanetworkopen.2024.38986.

[7] Stephen Gerry et al. "Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology". en. In: *BMJ* 369 (May 20, 2020), p. m1501. DOI: 10.1136/bmj.m1501.

[8] Christopher J Winslow et al. "The impact of a machine learning early warning score on hospital mortality: A multicenter clinical intervention trial". en. In: *Crit. Care Med.* 50 (9 Sept. 1, 2022), pp. 1339–1347. DOI: 10.1097/CCM.0000000000005492.

[9] Gabriel J Escobar et al. "Automated identification of adults at risk for in-hospital clinical deterioration". en. In: *N. Engl. J. Med.* 383 (20 Nov. 12, 2020), pp. 1951–1960. DOI: 10.1056/NEJMsa2001090.

[10] Brandon C Cummings et al. "External validation and comparison of a general ward deterioration index between diversely different health systems". en. In: *Crit. Care Med.* 51 (6 June 1, 2023), pp. 775–786. DOI: 10.1097/CCM.0000000000005837.

[11] Bryan Williams. "The National Early Warning Score: from concept to NHS implementation". en. In: *Clin. Med.* 22 (6 Nov. 2022), pp. 499–505. DOI: 10.7861/clinmed.2022-news-concept.

[12] C P Subbe et al. "Validation of a modified Early Warning Score in medical admissions". en. In: *QJM* 94 (10 Oct. 2001), pp. 521–526. DOI: 10.1093/qjmed/94.10.521.

[13] Ashish Vaswani et al. "Attention is all you need". In: *Adv. Neural Inf. Process. Syst.* 30 (2017).

[14] Pawel Renc et al. "Zero shot health trajectory prediction using transformer". en. In: *NPJ Digit. Med.* 7 (1 Sept. 19, 2024), p. 256. DOI: 10.1038/s41746-024-01235-0.

[15] Zhichao Yang et al. "TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records". en. In: *Nat. Commun.* 14 (1 Nov. 29, 2023), p. 7857. DOI: 10.1038/s41467-023-43715-z.

[16] Yikuan Li et al. "Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records". en. In: *IEEE J Biomed Health Inform* 27 (2 Feb. 2023), pp. 1106–1117. DOI: 10.1109/JBHI.2022.3224727.

[17] Xiaoliang Luo et al. "Large language models surpass human experts in predicting neuroscience results". en. In: *Nat. Hum. Behav.* (Nov. 27, 2024). DOI: 10.1038/s41562-024-02046-9.

[18] Arun James Thirunavukarasu et al. "Large language models in medicine". en. In: *Nat. Med.* 29 (8 Aug. 2023), pp. 1930–1940. DOI: 10.1038/s41591-023-02448-8.

[19] Zeljko Kraljevic et al. "Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study". en. In: *Lancet Digit Health* 6 (4 Apr. 2024), e281–e290. DOI: 10.1016/S2589-7500(24)00025-6.

[20] Matthew B A McDermott et al. "Event Stream GPT: A data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events". In: *Adv. Neural Inf. Process. Syst.* abs/2306.11547 (June 20, 2023). DOI: 10.48550/arXiv.2306.11547.

[21] Ethan Steinberg et al. "MOTOR: A time-to-event foundation model for structured medical records". In: *arXiv [cs.LG]* (Jan. 8, 2023).

[22] Yikuan Li et al. "BEHRT: Transformer for electronic health records". en. In: *Sci. Rep.* 10 (1 Apr. 28, 2020), p. 7155. DOI: 10.1038/s41598-020-62922-y.

[23] Hyewon Jeong et al. "Event-Based Contrastive Learning for medical time series". In: *arXiv [cs.LG]* (Dec. 15, 2023).

[24] Alistair E W Johnson et al. "MIMIC-IV, a freely accessible electronic health record dataset". en. In: *Sci Data* 10 (1 Jan. 3, 2023), p. 1. DOI: 10.1038/s41597-022-01899-x.

[25] Alistair Johnson et al. "Mimic-iv". In: *PhysioNet. Available online at: https://physionet. org/content/mimiciv/2.2/(accessed Oct 1, 2023)* (2023).

[26] A L Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". en. In: *Circulation* 101 (23 June 13, 2000), E215–20. DOI: 10.1161/01.cir.101.23.e215.

[27] *MEDS-DEV: Establishing Reproducibility and Comparability in Health AI*.

[28] Bert Arnrich et al. "Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health". In: *ICLR 2024 Workshop on Learning from Time Series For Health*. 2024.

[29] Feng Xie et al. "Benchmarking emergency department prediction models with machine learning and public electronic health records". en. In: *Sci. Data* 9 (1 Oct. 27, 2022), p. 658. DOI: 10.1038/s41597-022-01782-9.

[30] Gary B Smith et al. "The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death". en. In: *Resuscitation* 84 (4 Apr. 2013), pp. 465–470. DOI: 10.1016/j.resuscitation.2012.12.016.

[31] Sheng Zhang et al. "A multimodal biomedical foundation model trained from fifteen million image–text pairs". en. In: *NEJM AI* 2 (1 Jan. 2025). DOI: 10.1056/aioa2400640.

[32] T Olsson, A Terent, and L Lind. "Rapid Emergency Medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients". en. In: *J. Intern. Med.* 255 (5 May 2004), pp. 579–587. DOI: 10.1111/j.1365-2796.2004.01321.x.

[33] Matthew M Churpek et al. "Derivation of a cardiac arrest prediction model using ward vital signs". en. In: *Crit. Care Med.* 40 (7 July 2012), pp. 2102–2108. DOI: 10.1097/CCM.0b013e318250aa5a.

[34] David R Eitel et al. "The Emergency Severity Index triage algorithm version 2 is reliable and valid". en. In: *Acad. Emerg. Med.* 10 (10 Oct. 2003), pp. 1070–1080. DOI: 10.1197/s1069-6563(03)00350-6.

[35] Edward Choi et al. "Multi-layer representation learning for medical concepts". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco California USA). New York, NY, USA: ACM, Aug. 13, 2016. DOI: 10.1145/2939672.2939823.

[36] S Hochreiter and J Schmidhuber. "Long short-term memory". en. In: *Neural Comput.* 9 (8 Nov. 15, 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[37] Nassim Oufattole et al. "MEDS-Tab: Automated tabularization and baseline methods for MEDS datasets". In: *arXiv [cs.LG]* (Oct. 31, 2024).

[38] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA). KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 13, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. (Visited on 02/25/2024).

[39] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". en. In: *Nat. Methods* 17 (3 Mar. 2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[40] F Pedregosa et al. "Scikit-learn: Machine learning in python journal of machine learning research". In: *Journal of machine learning research* 12 (2011), pp. 2825–2830.

[41] Chuizheng Meng et al. "Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset". en. In: *Sci. Rep.* 12 (1 May 3, 2022), p. 7166. DOI: 10.1038/s41598-022-11012-2.

[42] Hugh Logan Ellis et al. "The early warning paradox". en. In: *NPJ Digit. Med.* 8 (1 Feb. 3, 2025), p. 81. DOI: 10.1038/s41746-024-01408-x.

[43] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. "Machine learning in medicine". en. In: *N. Engl. J. Med.* 380 (14 Apr. 4, 2019), pp. 1347–1358. DOI: 10.1056/NEJMra1814259.

Supplementary Materials

Table S1: **ETHOS performance on the ARES tasks with a breakdown for demographic subgroups.** This table presents the predictive performance (AUROC with 95% confidence intervals) of ETHOS (top) and MEDS-Tab (bottom) for four critical clinical outcomes used in ARES: Hospital Mortality, ICU Admission, Prolonged Hospital Stay (>10 days), and a Composite Risk Score (HM+IA+PS). The prevalence rates of each outcome are provided for reference. Performance metrics are further stratified by gender and race to assess potential disparities in model performance across demographic subgroups.

	Hospital Mortality	**ICU Admission**	**Prolonged Stay**	**Composite (HM+IA+PS)**
Prevalence (%)	1.85	15.44	9.01	20.39
		ETHOS		
Overall	0.936 [0.930, 0.944]	0.927 [0.927, 0.929]	0.849 [0.845, 0.848]	0.899 [0.898, 0.902]
Gender				
Female	0.939 [0.927, 0.945]	0.926 [0.918, 0.928]	0.852 [0.843, 0.862]	0.897 [0.894, 0.902]
Male	0.932 [0.921, 0.938]	0.927 [0.924, 0.931]	0.844 [0.836, 0.851]	0.900 [0.897, 0.906]
Race				
Asian	0.953 [0.947, 0.965]	0.939 [0.938, 0.955]	0.871 [0.854, 0.878]	0.913 [0.907, 0.926]
Black	0.959 [0.954, 0.970]	0.927 [0.926, 0.939]	0.867 [0.860, 0.877]	0.899 [0.888, 0.910]
Hispanic	0.942 [0.933, 0.981]	0.940 [0.934, 0.948]	0.872 [0.849, 0.883]	0.909 [0.893, 0.919]
Other	0.977 [0.962, 0.990]	0.961 [0.957, 0.969]	0.863 [0.837, 0.883]	0.927 [0.918, 0.940]
Unknown	0.890 [0.875, 0.910]	0.933 [0.921, 0.941]	0.796 [0.773, 0.799]	0.933 [0.927, 0.947]
White	0.920 [0.909, 0.925]	0.918 [0.915, 0.920]	0.840 [0.838, 0.843]	0.890 [0.888, 0.893]
		MEDS-Tab		
Overall	0.887 [0.882, 0.896]	0.918 [0.916, 0.919]	0.815 [0.810, 0.819]	0.879 [0.875, 0.879]
Gender				
Female	0.898 [0.892, 0.902]	0.916 [0.912, 0.919]	0.822 [0.817, 0.825]	0.877 [0.873, 0.881]
Male	0.876 [0.871, 0.884]	0.918 [0.913, 0.918]	0.807 [0.801, 0.808]	0.878 [0.872, 0.880]
Race				
Asian	0.895 [0.825, 0.909]	0.916 [0.906, 0.920]	0.819 [0.795, 0.847]	0.877 [0.860, 0.885]
Black	0.918 [0.892, 0.948]	0.920 [0.909, 0.928]	0.836 [0.824, 0.836]	0.874 [0.862, 0.883]
Hispanic	0.890 [0.846, 0.954]	0.922 [0.901, 0.927]	0.851 [0.839, 0.853]	0.900 [0.880, 0.907]
Other	0.933 [0.889, 0.945]	0.943 [0.938, 0.965]	0.830 [0.804, 0.854]	0.915 [0.902, 0.930]
Unknown	0.789 [0.779, 0.824]	0.953 [0.950, 0.958]	0.750 [0.705, 0.783]	0.926 [0.922, 0.942]
White	0.871 [0.868, 0.878]	0.907 [0.904, 0.909]	0.806 [0.798, 0.807]	0.867 [0.865, 0.867]

Table S2: **Demographic characteristics of the dataset analyzed in this study.** Performance Comparison of ETHOS and MEDS-Tab Across Clinical Outcomes and Demographic Subgroups. This table presents the predictive performance (AUROC with 95% confidence intervals) of ETHOS (top) and MEDS-Tab (bottom) for four critical clinical outcomes used in ARES: Hospital Mortality, ICU Admission, Prolonged Hospital Stay (>10 days), and a Composite Risk Score (HM+IA+PS). The prevalence rates of each outcome are provided for reference. Performance metrics are further stratified by gender and race to assess potential disparities in model performance across demographic subgroups.

	Train/Validation	**Test**	**Total**
Patient Number	269,741	29,971	299,712
Mean Age (Std.)	48.5 (20.9)	48.6 (20.9)	48.5 (20.9)
Gender (%)			
Female	142,696 (52.9)	15,857 (52.9)	158,553 (52.9)
Male	127,045 (47.1)	14,114 (47.1)	141,159 (47.1)
Race (%)			
Unknown	115,437 (42.8)	12,684 (42.3)	128,121 (42.7)
White	110,408 (40.9)	12,369 (41.3)	122,777 (41.0)
Black	21,410 (7.9)	2,321 (7.7)	23,731 (7.9)
Hispanic	9,214 (3.4)	1,023 (3.4)	10,237 (3.4)
Asian	6,802 (2.5)	787 (2.6)	7,589 (2.5)
Other	6,470 (2.4)	787 (2.6)	7,257 (2.4)
Marital Status (%)			
Unknown	114,234 (42.3)	12,603 (42.1)	126,837 (42.3)
Married	70,269 (26.1)	7,811 (26.1)	78,080 (26.1)
Single	60,915 (22.6)	6,793 (22.7)	67,708 (22.6)
Widowed	14,243 (5.3)	1,670 (5.6)	15,913 (5.3)
Divorced	10,080 (3.7)	1,094 (3.7)	11,174 (3.7)

Table S3: **Prediction of Hospitalization At Triage.** Performance comparison of various models for predicting hospitalization at triage, evaluated using AUROC, AUPRC, sensitivity, and specificity (95% confidence intervals in brackets). The thresholds for sensitivity and specificity were determined by finding the operating point on the ROC curve closest to (0,1). ETHOS demonstrates superior performance across all metrics, achieving the highest AUROC (0.912), AUPRC (0.887), sensitivity (0.849), and specificity (0.820), outperforming all other methods, including traditional scoring systems and machine learning models.

	AUROC	AUPRC	Sensitivity	Specificity
LR	0.809 [0.807, 0.813]	0.773 [0.770, 0.780]	0.751 [0.729, 0.757]	0.723 [0.720, 0.747]
RF	0.817 [0.814, 0.821]	0.785 [0.778, 0.791]	0.767 [0.735, 0.773]	0.716 [0.714, 0.748]
GB	0.819 [0.816, 0.822]	0.792 [0.787, 0.797]	0.753 [0.732, 0.771]	0.728 [0.712, 0.750]
MLP	0.823 [0.821, 0.827]	0.797 [0.792, 0.803]	0.748 [0.740, 0.778]	0.740 [0.720, 0.752]
ESI	0.712 [0.708, 0.716]	0.632 [0.628, 0.639]	0.584 [0.577, 0.591]	0.784 [0.782, 0.790]
NEWS	0.581 [0.577, 0.585]	0.555 [0.548, 0.559]	0.563 [0.554, 0.568]	0.546 [0.542, 0.552]
NEWS2	0.565 [0.561, 0.569]	0.538 [0.532, 0.543]	0.519 [0.510, 0.526]	0.570 [0.567, 0.576]
REMS	0.666 [0.661, 0.669]	0.605 [0.599, 0.610]	0.605 [0.553, 0.717]	0.641 [0.544, 0.712]
MEWS	0.558 [0.555, 0.561]	0.521 [0.516, 0.527]	0.296 [0.292, 0.299]	0.812 [0.807, 0.817]
CART	0.673 [0.668, 0.676]	0.617 [0.609, 0.622]	0.703 [0.699, 0.707]	0.578 [0.571, 0.583]
Med2Vec	0.815 [0.812, 0.818]	0.779 [0.774, 0.783]	0.741 [0.732, 0.756]	0.739 [0.726, 0.754]
AutoScore	0.794 [0.791, 0.797]	0.755 [0.750, 0.761]	0.745 [0.714, 0.749]	0.698 [0.692, 0.730]
MEDS-Tab	0.846 [0.842, 0.850]	0.862 [0.859, 0.866]	0.723 [0.713, 0.729]	0.807 [0.801, 0.819]
ETHOS (ours)	0.912 [0.909, 0.915]	0.887 [0.881, 0.890]	0.849 [0.846, 0.852]	0.820 [0.817, 0.823]

Table S4: **Prediction of Critical Outcome Within 12h At Triage.** Performance comparison of various models for predicting critical outcomes within 12 hours of triage, evaluated using AUROC, AUPRC, sensitivity, and specificity (95% confidence intervals in brackets). The thresholds for sensitivity and specificity were determined by finding the operating point on the ROC curve closest to (0,1). ETHOS achieves the highest performance across most of the metrics, with an AUROC of 0.937, AUPRC of 0.649, sensitivity of 0.858, and specificity of 0.863, substantially outperforming all other methods, including traditional scoring systems and machine learning models.

	AUROC	AUPRC	Sensitivity	Specificity
LR	0.869 [0.864, 0.872]	0.327 [0.309, 0.338]	0.804 [0.795, 0.822]	0.777 [0.763, 0.784]
RF	0.875 [0.868, 0.881]	0.385 [0.368, 0.405]	0.813 [0.789, 0.825]	0.785 [0.783, 0.809]
GB	0.884 [0.880, 0.887]	0.404 [0.386, 0.415]	0.804 [0.799, 0.829]	0.799 [0.774, 0.800]
MLP	0.888 [0.883, 0.891]	0.397 [0.378, 0.407]	0.817 [0.803, 0.838]	0.795 [0.782, 0.812]
ESI	0.807 [0.802, 0.812]	0.198 [0.188, 0.207]	0.877 [0.870, 0.888]	0.640 [0.637, 0.644]
NEWS	0.638 [0.625, 0.654]	0.155 [0.144, 0.165]	0.458 [0.439, 0.480]	0.798 [0.796, 0.803]
NEWS2	0.621 [0.609, 0.635]	0.141 [0.132, 0.149]	0.596 [0.404, 0.612]	0.536 [0.535, 0.826]
REMS	0.679 [0.668, 0.689]	0.109 [0.102, 0.117]	0.674 [0.656, 0.693]	0.605 [0.602, 0.608]
MEWS	0.621 [0.610, 0.631]	0.112 [0.106, 0.117]	0.442 [0.421, 0.461]	0.774 [0.770, 0.778]
CART	0.706 [0.695, 0.713]	0.154 [0.142, 0.161]	0.591 [0.569, 0.603]	0.723 [0.719, 0.727]
Med2Vec	0.874 [0.870, 0.878]	0.334 [0.313, 0.347]	0.825 [0.794, 0.838]	0.767 [0.761, 0.797]
AutoScore	0.857 [0.853, 0.861]	0.311 [0.294, 0.319]	0.789 [0.753, 0.813]	0.762 [0.745, 0.794]
MEDS-Tab	0.814 [0.805, 0.822]	0.420 [0.403, 0.440]	0.776 [0.761, 0.790]	0.663 [0.658, 0.667]
ETHOS (ours)	0.937 [0.932, 0.943]	0.649 [0.628, 0.666]	0.858 [0.850, 0.867]	0.863 [0.856, 0.869]

Table S5: **Prediction of Emergency Department Re-presentation Within 72h.** Performance comparison of various models for predicting emergency department re-presentation within 72 hours, evaluated using AUROC, AUPRC, sensitivity, and specificity (95% confidence intervals in brackets). The thresholds for sensitivity and specificity were determined by finding the operating point on the ROC curve closest to (0,1). ETHOS demonstrates superior performance, achieving the highest AUROC (0.740), AUPRC (0.199), sensitivity (0.659), and specificity (0.696), outperforming all other methods and showcasing its effectiveness for this challenging task.

	AUROC	AUPRC	Sensitivity	Specificity
LR	0.677 [0.661, 0.693]	0.160 [0.138, 0.179]	0.571 [0.535, 0.645]	0.683 [0.626, 0.712]
RF	0.672 [0.654, 0.680]	0.147 [0.125, 0.158]	0.570 [0.552, 0.652]	0.688 [0.599, 0.690]
GB	0.698 [0.681, 0.713]	0.165 [0.147, 0.181]	0.616 [0.576, 0.673]	0.667 [0.609, 0.709]
MLP	0.694 [0.681, 0.706]	0.162 [0.144, 0.180]	0.628 [0.591, 0.659]	0.660 [0.615, 0.690]
Med2Vec	0.656 [0.646, 0.670]	0.133 [0.121, 0.150]	0.564 [0.547, 0.600]	0.663 [0.624, 0.679]
LSTM	0.689 [0.676, 0.702]	0.168 [0.147, 0.183]	0.654 [0.574, 0.661]	0.612 [0.609, 0.679]
AutoScore	0.624 [0.612, 0.636]	0.074 [0.068, 0.082]	0.589 [0.548, 0.635]	0.601 [0.563, 0.640]
MEDS-Tab	0.696 [0.679, 0.713]	0.163 [0.143, 0.185]	0.635 [0.598, 0.680]	0.661 [0.633, 0.704]
ETHOS (ours)	0.740 [0.723, 0.760]	0.199 [0.171, 0.230]	0.659 [0.643, 0.678]	0.696 [0.685, 0.708]

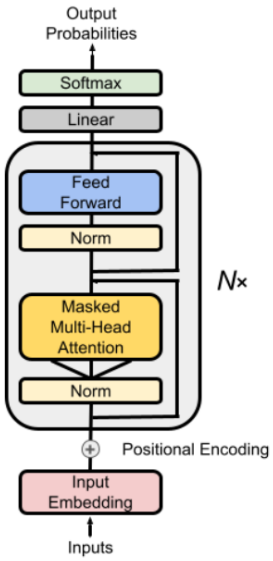
Table S6: **Summary of Token and Timeline Statistics.** This table presents a comprehensive overview of the token and timeline data in the training, test, and combined datasets. Key metrics include the total number of tokens and timelines, along with statistics on timeline lengths such as the longest timeline, median, mean, and shortest timeline. The number of unique timeline tokens is also reported. The final section breaks down the encoding of timeline tokens into categories, such as time intervals, quantiles, medications, diagnoses, procedures, laboratory results, vitals, and other clinical features. This summary highlights the diversity and complexity of the tokenized data used in the study.

	Train/Validation	**Test**	**Total**
Tokens	321,238,835	35,942,101	357,180,936
Timelines	257,082	28,540	285,622
Timeline Lengths			
Longest	221,122	106,936	221,122
Q3	1,041	1,052	1,042
Median	322	331	323
Mean	1,249	1,259	1,250
Q1	114	115	114
Shortest	2	2	2
Unique	13,094	4,940	13,631
Unique Timeline Tokens	4,495	3,947	4,495
Timeline Tokens Encoding			
Time Intervals	19	19	19
Quantiles	10	10	10
Medications	312	275	312
Diagnoses	2,989	2,542	2,989
Procedures	34	34	34
Labs	200	200	200
Vitals	6	6	6
HCPCS	66	37	66
Inpatient Stays	29	29	29
Emergency Department	7	7	7
DRGs	772	737	772
BMI	10	10	10

Table S7: **Overview of the data sources and their corresponding columns used in this work from the MIMIC-IV database and its extension MIMIC-IV-ED.** The table groups the data into three main categories: ED (Emergency Department), hosp (Hospital), and ICU (Intensive Care Unit). For each category, the associated tables and the specific columns extracted for the study are listed, highlighting key variables relevant to patient care and outcomes, such as identifiers (e.g., stay_id, hadm_id), timestamps (e.g., intime, charttime), and clinical observations (e.g., vitalsign, labresults). These selections were guided by the objectives of the study to comprehensively model patient trajectories and outcomes.

Data Source	Used Columns
ed	
diagnosis	icd_version, icd_code, stay_id
edstays	intime, arrival_transport, hadm_id stay_id, outtime, disposition hadm_id, stay_id
pyxis	name, charttime, stay_id
triage	acuity, stay_id
vitalsign	temperature, charttime, stay_id heartrate, charttime, stay_id resprate, charttime, stay_id o2sat, charttime, stay_id sbp, dbp, charttime stay_id, pain, charttime stay_id
hosp	
admissions	admission_type, admission_location, admittime insurance, marital_status, race hadm_id, discharge_location, dischtime hadm_id
diagnoses_icd	icd_version, icd_code, hadm_id
drgcodes	drg_type, drg_code, description hadm_id
emar	medication, event_txt, charttime hadm_id, emar_id, emar_seq
hcpcsevents	short_description, hadm_id, chartdate
labevents	itemid, valueuom, hadm_id charttime, valuenum, value
omr	result_name, result_value, chartdate
patients	gender, dod
procedures_icd	icd_version, icd_code, hadm_id chartdate
transfers	eventtype, careunit, intime hadm_id
icu	
icustays	first_careunit, intime, hadm_id stay_id, last_careunit, outtime hadm_id, stay_id

Figure S1: **Model Architecture and Hyperparameter Overview.** (Left) The architecture of the transformer-based model, following the standard GPT design, includes multiple layers of masked multi-head attention and feed-forward modules, normalized at each step and combined with positional encodings. (Right) Summary of the hyperparameters used for model training and their explored ranges. The final model uses 6 layers, a context size of 2048, an embedding size of 768, 12 attention heads, a dropout rate of 0.3, and a batch size of 32. Additional information includes the percentage of discarded ambiguous inference repetitions (0.2–0.3%) that appear when doing zero-shot inference.



Parameter name	Parameters used	Explored
Layer number	6	4-10
Context size	2048	512-4096
Embedding size	768	512-1024
Attention head number	12	8-16
Dropout	0.3	0-0.5
Batch size	32	16-64
Additional Information		
Model Type	GPT without bias	
Training Time	48h	
Inference repetitions	>=100	
Discarded ambiguous inference repetitions	0.2-0.3%	

Figure S2: **ROC Curves for ETHOS Across All Prediction Tasks.** ROC curves and corresponding area under the curve (AUC) values with 95% confidence intervals are shown for seven prediction tasks: Hospital Mortality, ICU Admission, Prolonged Stay (>10 days), Composite Outcome (Hospital Mortality + ICU Admission + Prolonged Stay), Hospitalization at Triage, Critical Outcome Within 12h at Triage, and Emergency Department (ED) Re-presentation Within 72h. Each plot includes the fitted ROC curve (orange), unique thresholds (crosses), and the 95% confidence interval (gray shading). ETHOS demonstrates high predictive performance across all tasks, with AUC values ranging from 0.740 (ED Re-presentation) to 0.936 (Hospital Mortality).

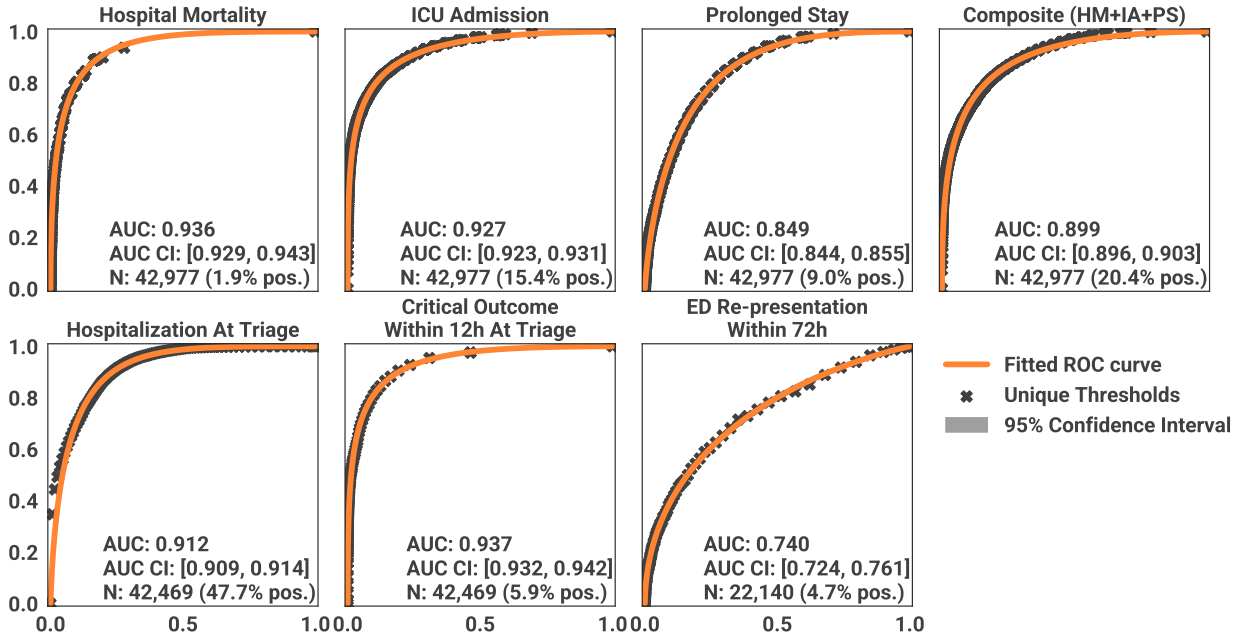


Figure S3: **Calibration Curves for ETHOS Predictions Across Clinical Outcomes with 95% Confidence Intervals Determined by Bootstrapping.** This figure presents calibration curves evaluating the reliability of ETHOS probability predictions across six key clinical outcomes: hospital mortality, ICU admission, prolonged hospital stay, composite risk score (HM+IA+PS), hospitalization at triage, critical outcome within 12 hours at triage, and ED re-presentation within 72 hours. The calibration curves compare predicted probabilities (x-axis) against observed event frequencies (y-axis), with perfect calibration represented by the dashed diagonal line, while the solid orange line shows ETHOS calibration performance, and the shaded gray region represents the 95% confidence interval (CI) derived from bootstrapping. Each plot includes the Brier score, a metric assessing probabilistic prediction accuracy, where lower values indicate better calibration, with 0.00–0.05 classified as excellent, 0.05–0.10 as good, 0.10–0.20 as acceptable, and values above 0.20 as poor calibration. ETHOS demonstrates excellent calibration for hospital mortality (Brier score: 0.014), critical outcome within 12 hours (0.031), and ED re-presentation (0.041), while ICU admission (0.064), prolonged stay (0.067), and the composite risk score (0.090) exhibit good calibration, closely following the ideal calibration curve. Hospitalization at triage (0.143) is categorized as acceptable calibration, with some deviations at higher predicted probabilities, suggesting areas for potential improvement. Overall, ETHOS exhibits strong calibration across most clinical tasks, particularly in predicting mortality, early critical deterioration, and ED re-presentation, with acceptable performance for hospitalization risk at triage. These findings highlight ETHOS's reliability in translating probability estimates into clinically meaningful risk stratifications, supporting its potential as a robust AI-driven decision support tool for real-time risk prediction and clinical decision-making.

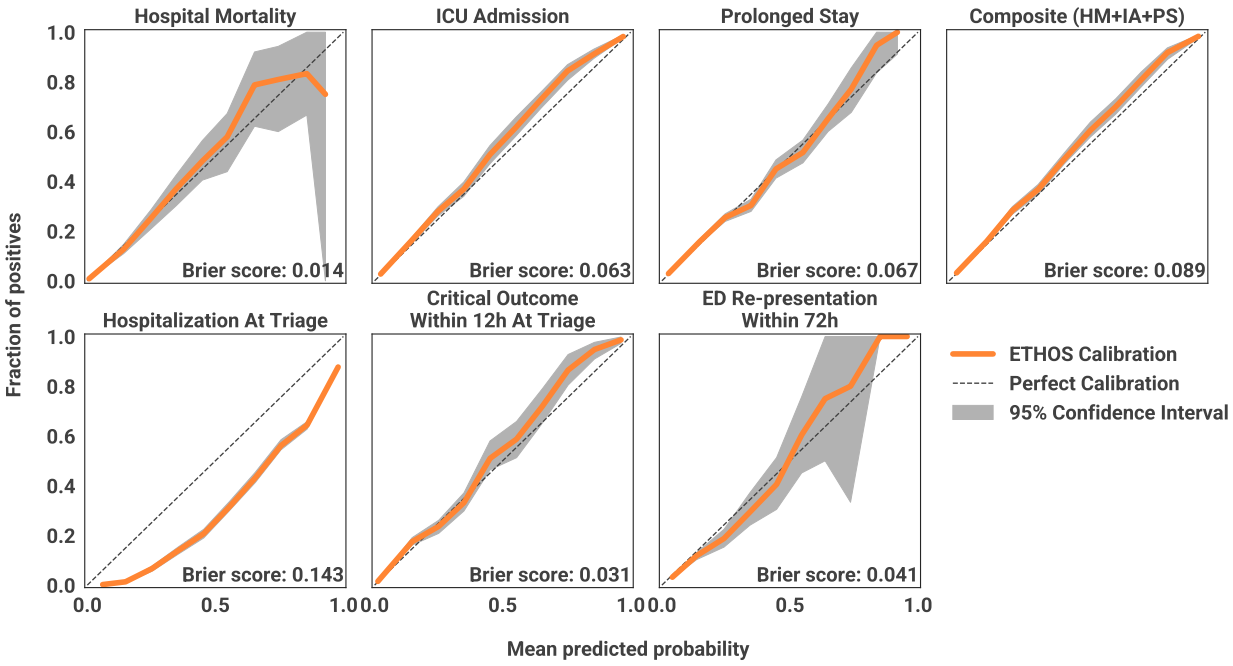
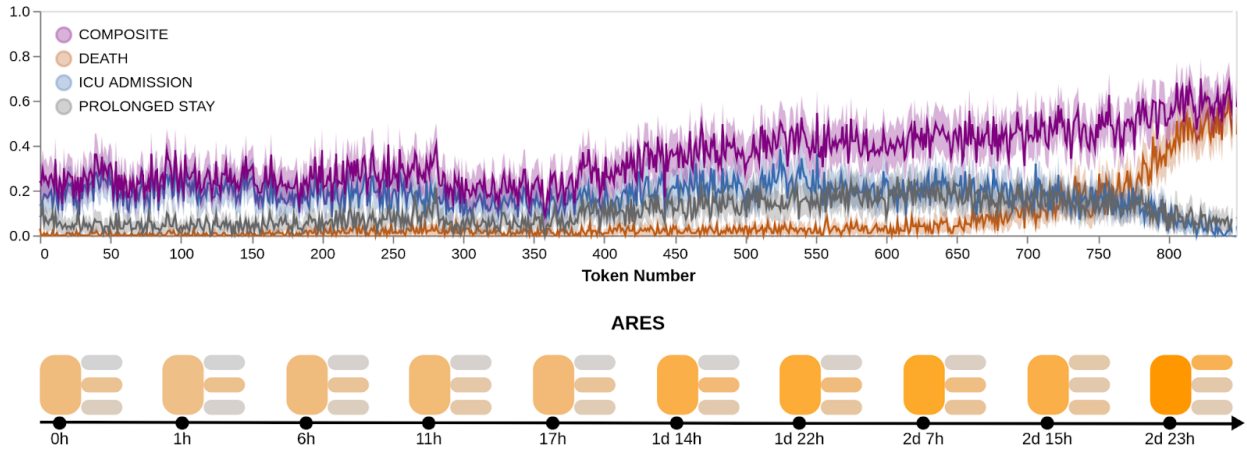
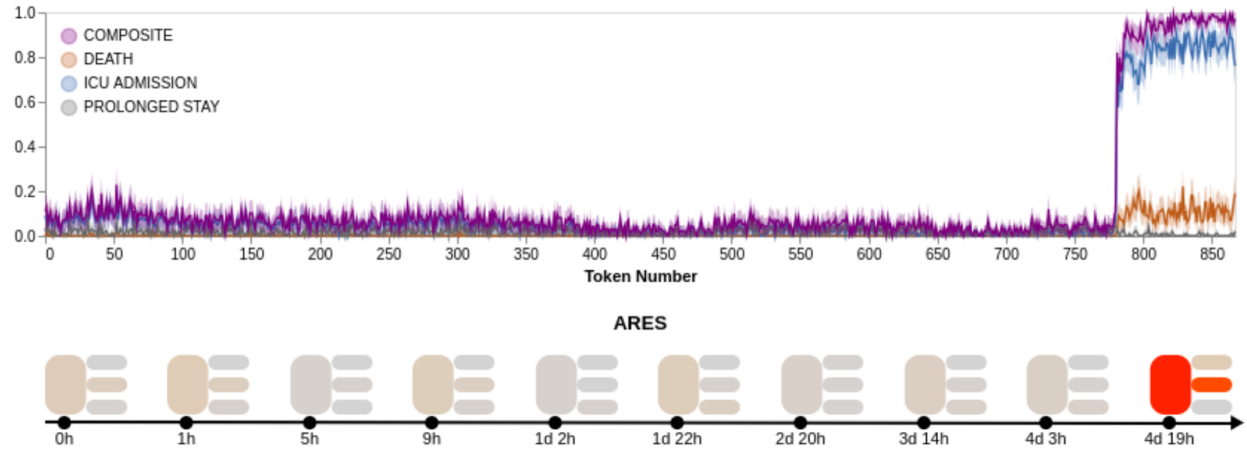


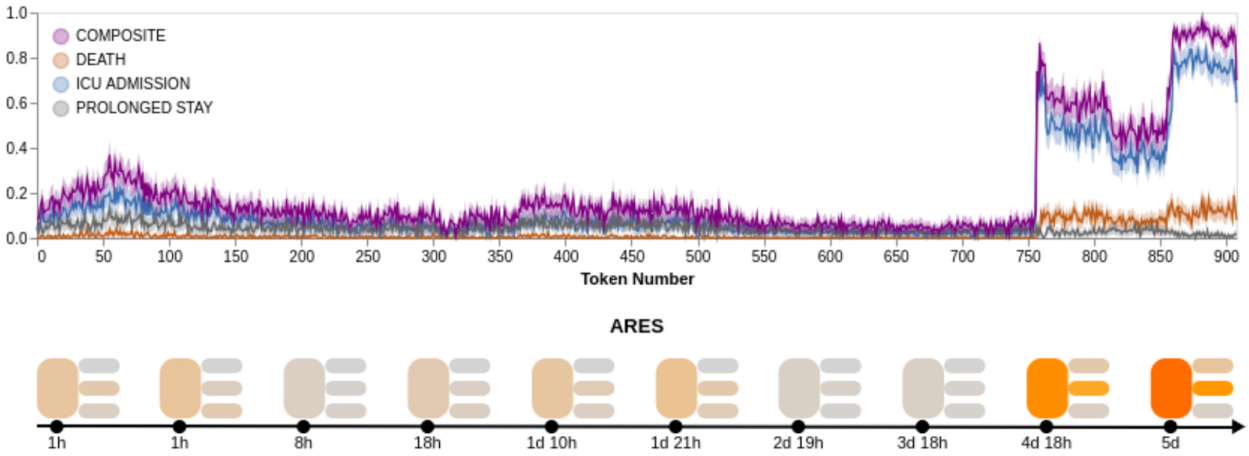
Table S8: **Risk Trajectories for Eight Patients from ED Presentation to Discharge, ICU Admission, or Death.**. This figure presents examples of risk trajectories for eight different patients, illustrating the dynamic evolution of risk predictions following presentation at the emergency department (ED). Each risk value is estimated from multiple (n=100) simulated fPHTs. The shaded area around each risk curve represents the 95% confidence interval (CI) for the predicted risk. The primary graphs plot risk progression as a function of the number of tokens generated since ED presentation, effectively modeling the temporal evolution of patient risk. The visualisation of ARES score is schematically represented below using 10 color-coded symbols corresponding to key risk categories (see Figures 1 and 2 in main paper). In some graphs, symbols corresponding to ICU admission risk are absent (e.g., E, F, G, and H) because these patients were already admitted to the ICU earlier, leading ARES to automatically exclude this risk component from consideration. The time axis under ARES represents actual elapsed time (in hours and days) since ED presentation. However, time progression on these axes is not linear, as the number of generated tokens does not directly correspond to real-time intervals. Instead, token generation occurs in discrete units determined by patient events. Notably, in case H, a sudden drop in prolonged stay risk occurs because ARES automatically reclassifies a risk of prolonged stay >10 days into prolonged stay >15 days, leading to an observed risk reduction. This drop is an inherent property of ARES modeling rather than a true change in patient status. All trajectories ultimately conclude when the patient either dies or is discharged.



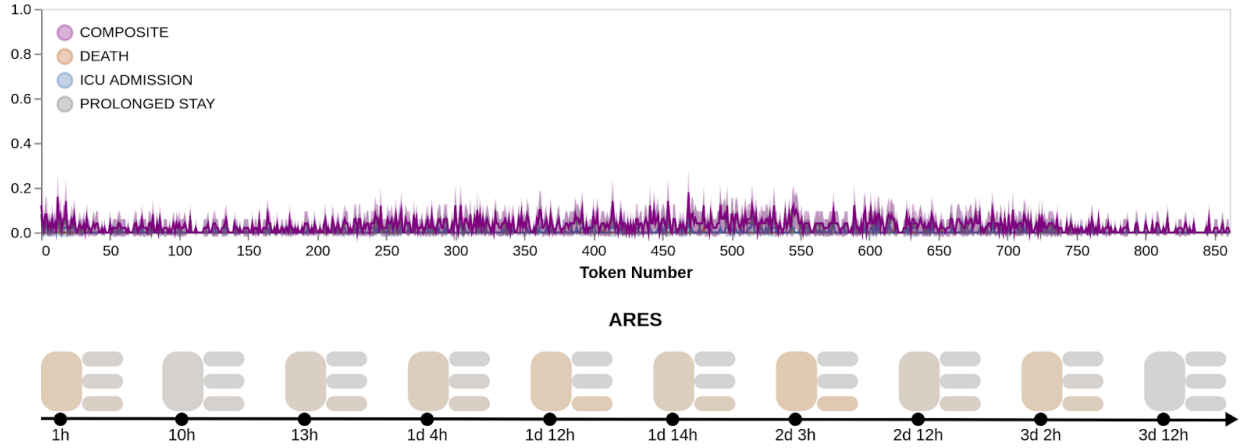
(A) Ends with death.



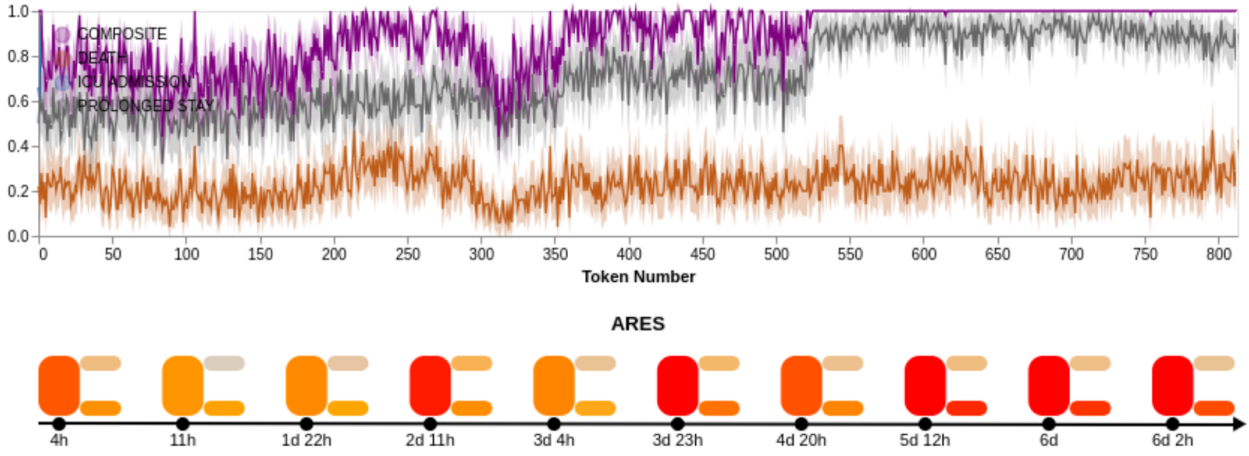
(B) Ends with ICU admission.



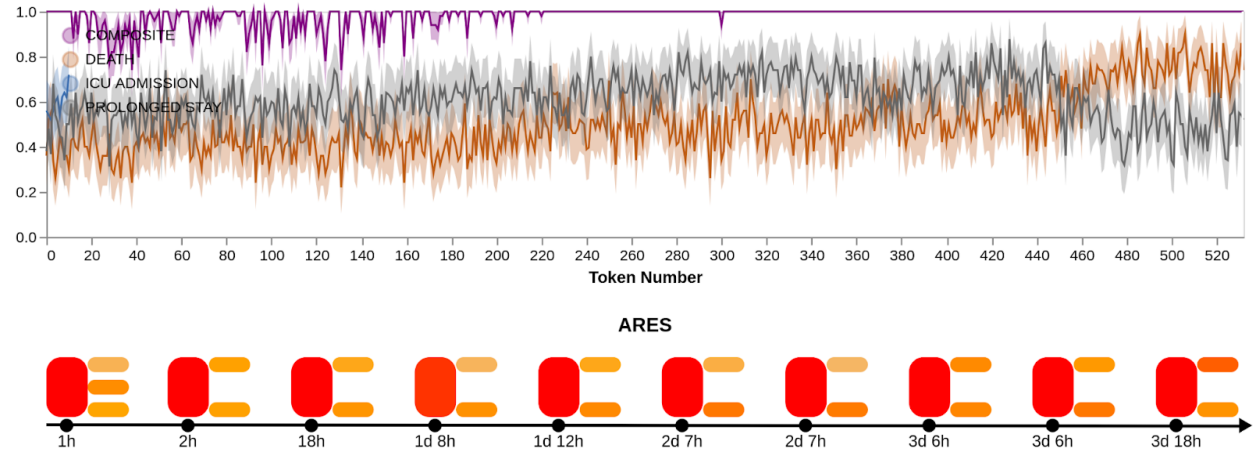
(C) Ends with ICU admission.



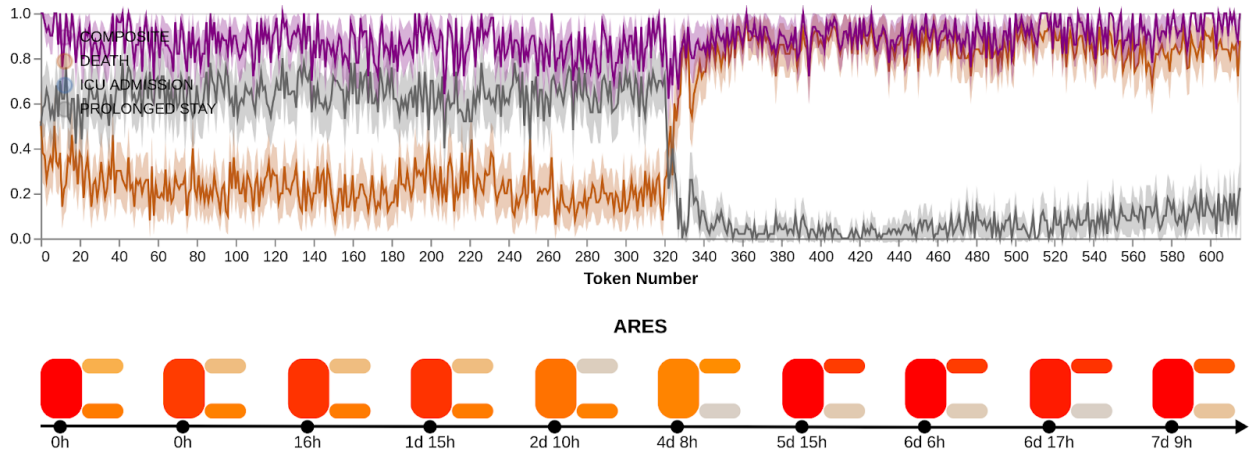
(D) Ends with hospital discharge.



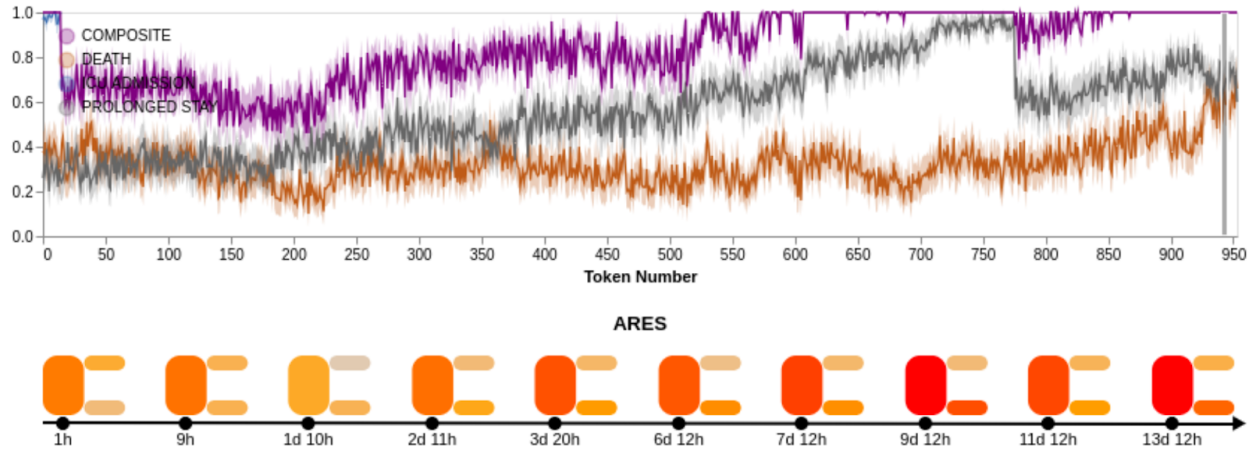
(E) Ends with death.



(F) Ends with death.



(G) Ends with death.



(H) Ends with death.

Table S9: **Detailed Token Statistics**. The table provides a detailed breakdown of the total number of tokens and unique tokens for each code group in the training, test, and combined datasets. Each code group represents a specific type of information, such as laboratory results (LAB), clinical classifications (e.g., ATC, ICD_CM), time intervals (e.g., 15m-45m, 12h-18h), and other key features like BMI, vitals, or discharge locations. The statistics summarize the diversity (#Unique) and frequency (Count) of tokens across datasets, offering insights into the distribution and variability of features used in the modeling process.

	Train		Test		Total	
	#Unique	Count	#Unique	Count	#Unique	Count
Code Group						
LAB	200	90,250,118	200	10,098,515	200	100,348,633
ATC	87	26,773,380	81	2,997,648	87	29,771,028
ATC_4	12	26,773,367	12	2,997,644	12	29,771,011
ATC_SFX	213	26,658,727	182	2,984,558	213	29,643,285
Q1	1	13,313,065	1	1,476,714	1	14,789,779
Q2	1	12,153,214	1	1,353,936	1	13,507,150
Q3	1	11,631,525	1	1,299,028	1	12,930,553
Q4	1	10,483,733	1	1,172,049	1	11,655,782
Q5	1	10,315,908	1	1,156,166	1	11,472,074
Q6	1	10,154,034	1	1,141,348	1	11,295,382
VITAL	6	9,946,752	6	1,113,072	6	11,059,824
Q7	1	9,574,210	1	1,076,334	1	10,650,544
ICD_CM	2,989	9,330,094	2,542	1,036,475	2,989	10,366,569
Q8	1	8,954,426	1	1,006,563	1	9,960,989
Q9	1	8,593,863	1	966,320	1	9,560,183
Q10	1	7,900,178	1	888,383	1	8,788,561
ICD_PCS	34	3,998,316	34	442,617	34	4,440,933
15m-45m	1	2,234,231	1	251,165	1	2,485,396
1h15m-2h	1	2,082,216	1	232,659	1	2,314,875
2h-3h	1	1,925,854	1	214,816	1	2,140,670
3h-5h	1	1,877,497	1	209,154	1	2,086,651
45m-1h15m	1	1,678,348	1	188,677	1	1,867,025
5m-15m	1	1,549,919	1	173,374	1	1,723,293
BMI	10	1,485,790	10	169,939	10	1,655,729
5h-8h	1	1,122,479	1	124,573	1	1,247,052
8h-12h	1	980,545	1	109,975	1	1,090,520
TRANSFER	38	750,441	38	83,393	38	833,834
12h-18h	1	708,241	1	79,051	1	787,292
2mt-6mt	1	465,225	1	52,313	1	517,538
=6mt	1	456,699	1	50,085	1	506,784
30d-2mt	1	430,807	1	48,696	1	479,503
12d-20d	1	388,256	1	44,259	1	432,515
DRG	772	388,255	737	42,977	772	431,232
HOSPITAL_DISCHARGE	1	388,254	1	42,977	1	431,231
DISCHARGE_LOCATION	10	388,254	10	42,977	10	431,231
INSURANCE	3	388,254	3	42,977	3	431,231
HOSPITAL_ADMISSION	1	388,254	1	42,977	1	431,231
ADMISSION_TYPE	3	388,254	3	42,977	3	431,231
ED_REGISTRATION	1	382,614	1	42,473	1	425,087
ED_OUT	1	382,614	1	42,473	1	425,087
ED_ACUITY	1	382,614	1	42,473	1	425,087
ED_TRANSPORT	4	382,614	4	42,473	4	425,087
20d-30d	1	340,809	1	38,149	1	378,958
4d-7d	1	333,877	1	38,375	1	372,252
7d-12d	1	328,988	1	37,916	1	366,904
1d-2d	1	307,351	1	34,627	1	341,978

Continued on next page

	Train		Test		Total	
Code Group	#Unique	Count	#Unique	Count	#Unique	Count
TIMELINE_END	1	257,082	1	28,540	1	285,622
2d-4d	1	227,549	1	25,932	1	253,481
18h-1d	1	225,224	1	25,242	1	250,466
HCPCS	66	127,052	37	13,731	66	140,783
ICU_ADMISSION	1	65,816	1	7,365	1	73,181
ICU_TYPE	9	65,816	9	7,365	9	73,181
ICU_DISCHARGE	1	65,816	1	7,365	1	73,181
SOFA	1	65,816	1	7,365	1	73,181
MEDS_DEATH	1	26,200	1	2,876	1	29,076

A Monte Carlo Justification for Probability Estimation

Let $p(\mathbf{x})$ denote the probability distribution over fPHTs as modeled by ETHOS where by $\mathbf{x}$ we indicate an fPHT. Suppose we want to estimate the probability of some event A regarding the future timeline. For instance, A could be the event "the patient death when admitted" or "the patient admitted to ICU." Formally,

$$\Pr(A) \;=\; \sum_{\mathbf{x}\in A} p(\mathbf{x}),$$

where the sum is over all sequences $\mathbf{x}$ for which the event A holds (i.e., $\mathbf{x} \in A$).

A. Monte Carlo Estimator

A straightforward Monte Carlo approach to approximate $\Pr(A)$ is as follows:

1. **Draw** N i.i.d. samples $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$ from the model $p(\mathbf{x})$.
2. **Define** an indicator function $I(\mathbf{x}^{(i)} \in A)$, which is 1 if the sample $\mathbf{x}^{(i)}$ lies in A, and 0 otherwise.
3. **Estimate** $\Pr(A)$ by the ratio

$$\hat{\Pr}(A) \;=\; \frac{1}{N}\sum_{i=1}^{N} I\big(\mathbf{x}^{(i)} \in A\big).$$

In other words, $\hat{\Pr}(A)$ is simply the fraction of samples whose corresponding timelines satisfy event A indicated as M/N in the text.

B. Unbiasedness

If the samples $\mathbf{x}^{(i)}$ are drawn exactly from $p(\mathbf{x})$, then for each sample,

$$\mathbb{E}[I(\mathbf{x}^{(i)} \in A)] \;=\; \Pr(\mathbf{x}^{(i)} \in A) \;=\; \Pr(A).$$

Hence,

$$\mathbb{E}\big[\hat{\Pr}(A)\big] \;=\; \mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^{N} I(\mathbf{x}^{(i)} \in A)\Big] \;=\; \Pr(A),$$

showing that $\hat{\Pr}(A)$ is an *unbiased* estimator of $\Pr(A)$.

C. Convergence by the Law of Large Numbers

By the Law of Large Numbers (LLN), as $N \to \infty$,

$$\hat{\Pr}(A) \;\xrightarrow{a.s.}\; \Pr(A),$$

meaning the simple ratio of "successes" (i.e., samples satisfying A) to total draws converges almost surely to the true probability.