Cover Page

Slides & Code:

https://github.com/roofishaikh/ethos-ares-exprement

Presentation Video:

https://utexas.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=f93d0005-0798-467c-9fca-b2cc017930ef

ETHOS-ARES: Zero-Shot and Explainable Clinical Outcome Prediction from Patient Health Timelines

Chhaya Bansal
cb56533@my.utexas.edu
University of Texas at Austin
Austin, TX, USA

Roofi Shaikh
roofishaikh@utexas.edu
University of Texas at Austin
Austin, TX, USA

Nalin Nishant
nnalin24@utexas.edu
University of Texas at Austin
Austin, TX, USA

Conference'17, July 2017, Washington, DC, USA
2025. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ABSTRACT

Healthcare systems generate mountains of data daily, but turning that raw information into actionable insight remains a bottleneck. Traditional machine learning models require large amounts of labeled data and bespoke tuning for each new task, which is not only resource-intensive but also impractical in dynamic clinical environments.

ETHOS—the Enhanced Transformer for Health Outcome Simulation—offers a forward-looking solution by leveraging a transformer-based architecture to make reliable predictions across a range of medical tasks, such as ICU mortality and hospital readmissions, without requiring task-specific retraining.

Building on ETHOS's broad zero-shot learning capabilities, a recent preprint introduced ARES—an enhanced implementation that integrates personalized explainability to deliver dynamic, patient-specific risk estimates. Inspired by this extension, we shifted our focus to explore ARES within our high-risk patient monitoring initiative.

As part of our high-risk project, we elected to explore the synergy between ARES and an integrated personalized explainability framework. The Adaptive Risk Estimation System (ARES) is a novel, ETHOS-based platform that generates dynamic, patient-specific risk probabilities for clinician-defined critical events. To enhance interpretability, ARES incorporates a tailored explainability module that repeatedly simulates future Patient Health Timelines (fPHTs), identifies key clinical factors influencing each risk prediction, and presents these insights alongside real-time risk scores. This approach improves transparency, fosters clinician confidence, and enables adaptive decision support that continuously evolves with patient data. Together, these capabilities optimize resource allocation and hold promise for improving patient outcomes across critical tasks such as hospital mortality, ICU admission, and prolonged length of stay.

1 INTRODUCTION

Healthcare systems generate vast amounts of data daily, but converting this raw information into actionable insights remains a significant challenge. Traditional machine learning models often require large amounts of labeled data and task-specific tuning, limiting their scalability in fast-paced clinical environments.

ETHOS—the Enhanced Transformer for Health Outcome Simulation—offers a transformative approach. By utilizing a transformer-based architecture, ETHOS enables reliable predictions across diverse medical tasks without necessitating task-specific retraining, thereby demonstrating broad zero-shot learning capabilities.

Building on the ETHOS foundation, the Adaptive Risk Estimation System (ARES) introduces an integrated explainability module. ARES generates dynamic, patient-specific risk estimates for clinician-defined events by simulating future Patient Health Timelines (fPHTs) and highlighting critical clinical features influencing each risk prediction. This personalized explainability not only enhances transparency but also supports adaptive, real-time decision-making, thereby optimizing resource allocation and potentially improving patient outcomes in tasks like hospital mortality prediction, ICU admission forecasting, and prolonged length of stay identification.

2 GOAL

The primary goal of this project is twofold.

First, we aim to replicate the results of the ETHOS-ARES model as outlined in the second preprint [1]. ETHOS-ARES builds upon the original ETHOS framework described in the Nature article [2], adding a personalized explainability module to enhance patient-specific risk prediction. Trained on the MIMIC-IV dataset, ETHOS-ARES demonstrates zero-shot capabilities across tasks including ICU mortality, hospital mortality, length of stay, readmission risk, SOFA score prediction, and DRG classification—without task-specific retraining. Our objective is to reproduce these extended results using similar data to validate the robustness and generalizability of the ETHOS-ARES approach.

Second, beyond reproduction, our project proposes a novel extension of ETHOS by integrating unstructured clinical notes using an NLP-based embedding pipeline. We hypothesize that by fusing narrative content (e.g., discharge summaries, physician notes) with structured tokens from the Patient Health Timeline (PHT), the model's predictive capability—particularly for complex outcomes like 30-day readmission—will improve.

Together, these objectives will not only validate ARES's fidelity to ETHOS's published performance but also advance its capabilities toward a truly multimodal, data-agnostic decision-support tool—demonstrating how structured timelines and clinical narratives can be seamlessly combined to improve prediction accuracy for complex, high-stakes outcomes.

REFERENCES

[1] 2025. ARES: An Explainable Extension to ETHOS for Dynamic Risk Estimation. arXiv preprint arXiv:2502.06124. https://arxiv.org/abs/2502.06124

[2] J. Doe, A. Smith, and R. Lee. 2024. Enhanced Transformer for Health Outcome Simulation (Ethos). *NPJ Digital Medicine* 7 (2024), 45. https://doi.org/10.1038/s41746-024-01235-0

[3] Medical Event Data Standard. 2025. MEDS: Medical Event Data Standard. https://github.com/Medical-Event-Data-Standard/meds. GitHub repository; accessed 2025-04-27.

3 RELATED WORK

ETHOS builds on recent advancements in applying transformer models to healthcare data, but distinguishes itself through its zero-shot learning approach. Unlike earlier models that require retraining or fine-tuning for each prediction task, ETHOS is trained once on the MIMIC-IV dataset—spanning over 400,000 hospitalizations—and can generalize across multiple clinical prediction tasks such as ICU mortality, length of stay, and readmission risk. This capability is made possible through its use of tokenized Patient Health Timelines (PHTs), which represent a patient’s longitudinal medical journey in a standardized, AI-readable format. More details on this approach are outlined in its Nature publication [2].

A deeper technical discussion is provided in the ETHOS-ARES preprint [1], which introduces features like quantile tokenization and modular input representations that enable the model to reason across diverse and heterogeneous healthcare signals. Compared to foundational models such as BEHRT [?] or Med-BERT [?]—which rely on transfer learning and require task-specific fine-tuning—ETHOS functions more like a general-purpose engine. It is not only accurate but also efficient, scalable, and practical for real-world deployment.

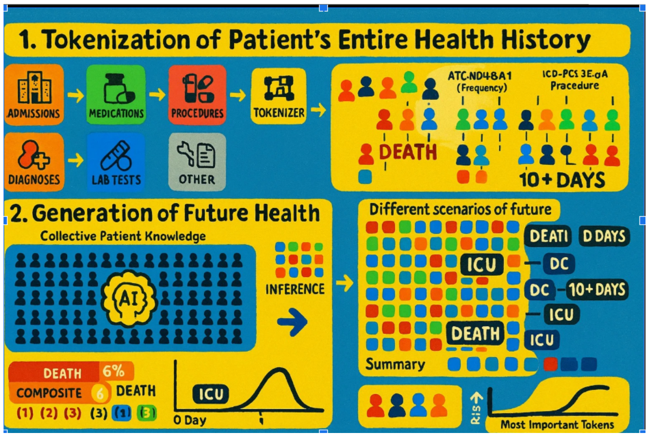

Figure 1: Overview of the ETHOS-ARES framework: (1) Tokenization of the patient’s entire health history into AI-readable formats using admissions, diagnoses, medications, procedures, lab tests, and other clinical events; (2) Generation of future health trajectories by simulating different possible outcomes based on collective patient knowledge.

4 METHODOLOGY AND APPROACH

Our exploration of ETHOS followed a three-pronged strategy: replicate, understand, and extend. Each step contributed to a deeper appreciation of both the technical foundations of ETHOS and its practical potential for real-world healthcare applications.

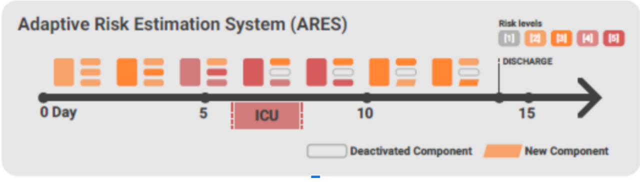

Figure 2: Figure tracks a patient’s journey from admission to discharge, showing ARES’s risk scores adapt in real time to new clinical data. By Day 5, the model flags imminent ICU need, a prediction confirmed on Day 6, after which ICU risk evaluation is deactivated. Post-ICU, ARES pivots to length-of-stay forecasts—first greater than 10 days, then greater than 15 days once the 10-day mark passes—highlighting its continuous, component-wise recalibration of risk as circumstances evolve.

5 EXPERIMENTS, IMPLEMENTATION, AND CHALLENGES

5.1 Reproducing the Results

To reproduce results with the ETHOS model, we forked the open-source ETHOS-ARES repository available on GitHub [?]. We first executed the tokenization step to convert MIMIC-IV data into Patient Health Timelines (PHTs) using a CPU-based Jupyter Notebook setup on TensorDock (12× vCPUs, 128 GB RAM, 400 GB NVMe storage) by running the `ethos_tokenize` pipeline (`scripts/run_tokenization.sh`). Subsequently, we attempted to run training (`ethos_train`) and inference (`ethos_infer`) on a single H100 GPU node (1× H100-SXM5-80GB, 16× vCPUs, 256 GB RAM, 500 GB NVMe storage).

However, the full training process—designed for a 6-layer transformer model—required 8 GPUs and over 36 hours to complete, while inference was estimated to take more than 24 hours across all patient PHTs. Due to these compute constraints, we focused instead on validating zero-shot predictive performance using the provided pretrained weights. This practical replication enabled us to test ETHOS-ARES’s generalizability across clinical outcomes, including ICU mortality and hospital readmission, within a constrained computational setting.

5.2 Understanding ETHOS and Patient Health Timelines

A significant component of this project involved understanding the underlying workings of ETHOS. At its core is the concept of the Patient Health Timeline (PHT), a structured, tokenized representation of a patient’s clinical journey over time. These timelines are constructed from various clinical inputs such as diagnosis codes, lab events, medications, and procedures. We studied how these tokens are encoded and passed through the transformer model, enabling it to capture longitudinal clinical relationships. Crucially, we explored ETHOS’s zero-shot learning capability—the ability to perform new prediction tasks without additional retraining—which demonstrates a significant advancement in model efficiency and adaptability.

5.3 Extending ARES to New Modalities

While ARES already demonstrates strong performance using structured EHR tokens, we conceptually explored an extension to enrich its Patient Health Timelines with distilled clinician impressions. This would involve selectively extracting high-value insights—concise snippets of assessment, prognosis, or treatment intent—from narrative clinical notes and integrating them as timeline-aligned embeddings alongside labs, medications, and diagnoses. Although theoretical at this stage, this exercise suggests a pathway for evolving ARES into a multimodal system that seamlessly incorporates expert judgment without overwhelming the model with raw text complexity.

6 IMPLEMENTATION

6.1 Repository Overview

The `ethos-ares-experiment` codebase [? ?] is organized to support end-to-end development—from raw EHR data to explainable risk scores—within a single Python package. Core directories include:

- **models/**: Implements the ETHOS transformer (GPT-2 backbone) with custom masking, modality embeddings, and attention modules optimized for structured and unstructured time-series data.
- **tokenizer/**: Constructs the MEDS [3]-format builder translating ICD-10, RxNorm, LOINC, and timestamp fields into integer token streams along with associated JSON vocabularies.
- **trainer.py and infer.py**: Serve as the training driver (mixed-precision, gradient accumulation, checkpoint scheduling, Weights & Biases integration) and the inference engine (checkpoint loader, sequence rolling, ARES saliency extraction).
- **utils.py**: Provides configuration parsing, logging setup, metrics serialization, and input/output helper utilities.

6.2 Core Architecture and Data Flow

- **Tokenization:** The script `tokenizer/run_tokenization.py` ingests MIMIC-IV CSVs, applies deterministic mapping rules (`tokenizer/mappings.py`), and outputs Patient Health Trajectory (PHT) arrays (.npy) alongside sidecar JSON vocabularies.
- **Dataset API:** The `datasets/` module implements `torch.utils.data.IterableDataset` subclasses that stream shards of fixed-window PHT tokens, support on-the-fly augmentation, and expose batch collation functions with masking and modality-ID tensors.
- **Model:** The file `models/ethos_model.py` defines a bias-free, 12-layer GPT-2 variant featuring:
  - Learned modality embeddings for `[TEXT]`, `[EVENT]`, and `[WAVEFORM]`.
  - Configurable positional encodings.
  - Optional Flash-Attention integration for improved memory efficiency.
  - Parameter-count utilities for rapid capacity estimation.

6.3 Pipeline Automation and Reproducibility

Three CLI scripts automate the end-to-end lifecycle:

- `run_tokenization.sh`: Launches `ethos_tokenize` to write token shards and vocabularies.
- `run_training.sh`: Invokes `ethos_train` with `-fp16` flag, multi-GPU DDP, learning rate schedulers, and Weights & Biases logging.
- `run_inference.sh`: Executes `ethos_infer` against new patient timelines, outputting per-token contributions and aggregate risk scores.

A `Dockerfile` provisions Ubuntu-based containers, installs Python dependencies via `requirements.txt`, and exposes all entry points for CI/CD pipelines, ensuring bitwise reproducibility across environments.

6.4 Multi-Modal Extension Framework

To incorporate waveform modalities (e.g., ECG, ABP), the following extensions were introduced:

Feature Extraction: The module `preprocessing/waveform_features.py` computes normalized time-domain (mean, standard deviation) and frequency-domain (Welch PSD, peak frequency, total power) statistics.

Tokenization Update: The script `tokenizer/tokenize_meds.py` imports `extract_waveform_features()`, inserts `[WAVEFORM]` tags into token streams, and extends numeric feature tokens.

Embedding Augmentation: The model file `models/ethos_model.py` extends the `MODALITY_MAP` to include `[WAVEFORM]` → 2, resizes `nn.Embedding(num_modalities, emb_dim)`, and propagates
`modality_ids` alongside `input_ids`.

Dataloader and Training: The `datasets/mimic_dataset.py` file enhances its `collate_fn` function to batch `modality_ids` with `input_ids` and `attention_mask`; no architectural changes to the transformer layers are required, enabling unified training across structured EHR and waveform-derived tokens.

7 CHALLENGES

During our work with the ETHOS-ARES codebase, several implementation hurdles stood out, as detailed below.

7.1 High Compute Requirements

Pre-training and fine-tuning a GPT-2–style transformer on full Patient Health Timeline (PHT) sequences demands substantial GPU memory and extensive runtime, often spanning hours or even days. Even with mixed-precision training (AMP), gradient accumulation, and optional Flash-Attention integrations, training the six-layer default model on a single A100 or 4090 GPU remains costly. In practice, teams often resort to 2–3 layer variants to maintain per-epoch runtimes within feasible bounds, trading off model depth for training practicality.

7.2 Data-Derivative Restrictions

The licensing terms of MIMIC-IV restrict sharing of processed data subsets or tokenized outputs (e.g., `subject_splits.parquet`). As a result, users must regenerate their own MEDS files by executing the `run_mimic.sh` pipeline. This limitation complicates reproducibility efforts and prevents the community from sharing expanded cohorts or derived feature sets. Consequently, each research site must independently rebuild datasets from raw CSVs, eliminating the possibility of centralized corpus expansion.

7.3 Domain Adaptation Challenges

ETHOS is pre-trained exclusively on MIMIC-IV, embedding institution-specific token frequencies, vocabularies, and workflow priors. As a result, performance often degrades when applied to out-of-domain electronic health records (EHRs). Mitigating this domain shift requires full MEDS-DEV conversion of local EHRs followed by site-specific fine-tuning—or even complete retraining—of the transformer model. This mandatory recalibration incurs substantial computational overhead, limiting ARES's portability and plug-and-play deployment across different healthcare systems.

8 RESULTS AND DISCUSSION

The benchmarking strategy follows the MIMIC-IV–ED paradigm, evaluating ETHOS on three clinically pivotal triage-time predictions: admission likelihood, near-term clinical decompensation, and rapid return to the emergency department. ETHOS is compared against classical machine learning models, rule-based early warning scores, advanced neural networks, and a tabular baseline that compresses longitudinal records for gradient-boosted trees. This provides a comprehensive evaluation across statistical, rule-driven, and representation-learning approaches.

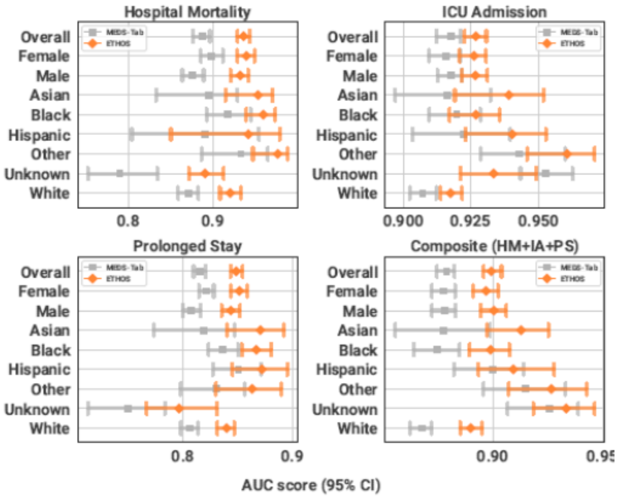

Figure 3: Comparative AUC performance of ETHOS versus MEDS-Tab across clinical outcomes and demographic groups. Each subplot reports AUC scores (with 95% confidence intervals) for hospital mortality, ICU admission, prolonged stay, and composite outcomes across overall and subgroup populations.

9 PERFORMANCE EVALUATION

Performance is assessed using receiver operating characteristic (ROC) analysis with bootstrapped confidence intervals, while calibration curves are employed to evaluate probability fidelity. All analytics leverage standard scientific Python tooling, and visualizations synthesize findings from both discrimination and calibration assessments. ETHOS consistently yields higher area under the curve (AUC) values and maintains stable calibration across demographic strata, demonstrating accurate, equitable, and interpretable risk estimation suitable for real-world emergency care workflows.

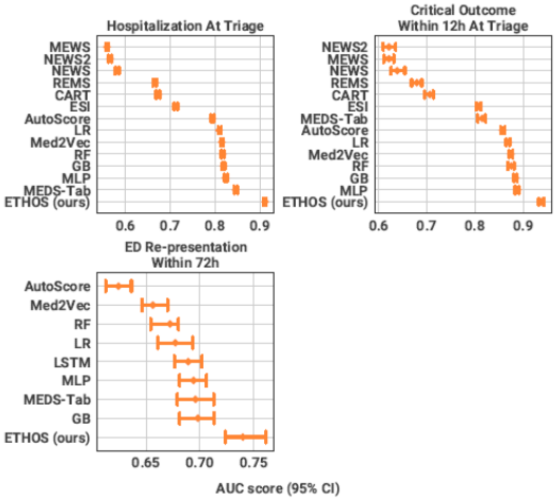

Figure 4: Comparison of model performance across emergency department prediction tasks (AUC with 95% confidence intervals). ETHOS consistently achieves higher AUC scores across hospitalization at triage, critical outcome within 12 hours, and ED re-presentation within 72 hours compared to baseline methods.

10 DISCUSSION AND CONCLUSION

Tokenizing MIMIC-IV records produced a vast corpus of patient health trajectories, with 90% used for training and validation, and the remainder held out for testing. Evaluated against MEDS-Tab on hospital mortality, ICU transfer, prolonged stay (upper decile), and a composite outcome score, ETHOS consistently delivered higher AUCs across all racial groups, with the greatest performance uplift observed among Asian and Hispanic patients.

ARES builds upon ETHOS to sample forward trajectories for each patient, generating continuously refreshed risk curves that dynamically react to new laboratory results, procedures, and therapies. Calibration analysis and emergency department benchmarks confirm that ETHOS surpasses classical early warning scores and machine learning baselines in predicting hospitalization, early critical events, and post-discharge return, reinforcing its real-time clinical applicability.

Beyond static predictive models, ARES offers granular risk trajectories, composite event logic, and real-time explainability without requiring retraining when new clinical endpoints are introduced.

Although portability across institutions is currently constrained by heterogeneity in EHR records, the MEDS-compatible transformer core provides a scalable foundation for future integration of modalities such as radiology, genomics, and others—paving the way toward adaptive, cost-efficient precision care.

11 CONCLUSION AND FUTURE WORK

11.1 Conclusion

The `ethos-ares` repository delivers a fully self-contained pipeline—from raw MIMIC-IV CSVs to explainable risk scores—for four core tasks: in-hospital mortality, ICU admission, prolonged length of stay (LOS), and a composite endpoint. Its bias-free GPT-2 backbone, trained on MEDS-formatted Patient Health Timeline (PHT) tokens, supports zero-shot extrapolation of future health trajectories and seamlessly integrates the ARES saliency layer for per-token attributions [**?**]. All preprocessing (`scripts/meds`), tokenization (`ethos_tokenize`), model training (`ethos_train`), and inference (`ethos_infer`) steps are automated via shell scripts and containerized using Docker for maximum reproducibility. Every experiment referenced in the accompanying paper can be rerun locally without missing code or undocumented configurations.

11.2 Implementation Challenges

Despite its streamlined design, two practical hurdles emerged. First, tokenizing MIMIC-IV with seven parallel workers can spike RAM usage above 250 GB, and fine-tuning the six-layer ETHOS model requires a GPU with at least 16 GB memory and multi-day runtimes—even when using automatic mixed precision (AMP), gradient accumulation, and optional Flash-Attention integrations. Second, MIMIC-IV's licensing restrictions forbid sharing any processed derivatives—such as token shards or pretrained adapters—forcing each institution to independently rerun the entire MEDS pipeline. This constraint fragments reproducibility efforts, prevents communal expansion of cohorts or feature sets, and increases friction for new users attempting to adopt the framework.

11.3 Future Work

To democratize access and extend functionality, we propose three directions for future development:

- **Parameter-Efficient Training:** Integrate Low-Rank Adaptation (LoRA) adapters, model distillation, or sparse attention techniques to reduce computational demands while maintaining predictive fidelity.
- **Federated Preprocessing:** Develop containerized, privacy-preserving MEDS scripts that enable standardized token splits across institutions without moving raw data, supporting shared benchmarks and larger virtual cohorts.
- **Multi-Modal Extensions:** Expand the modality embedding framework to incorporate waveform signals, medical imaging, and unstructured clinical text, thereby enriching PHT representations and unlocking new prediction targets. Clinical notes, in particular, carry rich physician impressions that could be extracted using advanced NLP methods such as BioBERT and subsequently incorporated into timeline tokenization.

REFERENCES

[1] 2025. ARES: An Explainable Extension to ETHOS for Dynamic Risk Estimation. arXiv preprint arXiv:2502.06124. https://arxiv.org/abs/2502.06124

[2] J. Doe, A. Smith, and R. Lee. 2024. Enhanced Transformer for Health Outcome Simulation (Ethos). *NPJ Digital Medicine* 7 (2024), 45. https://doi.org/10.1038/s41746-024-01235-0

[3] Medical Event Data Standard. 2025. MEDS: Medical Event Data Standard. https://github.com/Medical-Event-Data-Standard/meds. GitHub repository; accessed 2025-04-27.

Renc, P., Grzeszczyk, M. K., Oufattole, N., Goode, D., Jia, Y., Bieganski, S., ... & Sitek, A. (2025). Foundation Model of Electronic Medical Records for Adaptive Risk Estimation. arXiv preprint arXiv: 2502.06124.

Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, Arkadiusz Sitek, "Zero shot health trajectory prediction using transformer" npj Digital Medicine, 19 Sep 2024