A Transformer-Based Model for Zero-Shot Health Trajectory Prediction

Pawel Renc[1,2,3], Yugang Jia[4], Anthony E. Samir[1,2], Jaroslaw Was[3], Quanzheng Li[1,2], David W. Bates[5,2.6], Arkadiusz Sitek[1,2,*]

1.Massachusetts General Hospital, 2.Harvard Medical School, 3.AGH University of Science and Technology, 4.Massachusetts Institute of Technology, 5.Brigham and Women's Hospital, 6.Harvard Chan School of Public Health

*asitek@mgh.harvard.edu

Abstract: Integrating modern machine learning and clinical decision-making has great promise for mitigating healthcare's increasing cost and complexity. We introduce the Enhanced Transformer for Health Outcome Simulation (ETHOS), a novel application of the transformer deep-learning architecture for analyzing high-dimensional, heterogeneous, and episodic health data. ETHOS is trained using Patient Health Timelines (PHTs)—detailed, tokenized records of health events—to predict future health trajectories, leveraging a zero-shot learning approach.

ETHOS represents a significant advancement in foundation model development for healthcare analytics, eliminating the need for labeled data and model fine-tuning. Its ability to simulate various treatment pathways and consider patient-specific factors positions ETHOS as a tool for care optimization and addressing biases in healthcare delivery. Future developments will expand ETHOS' capabilities to incorporate a wider range of data types and data sources. Our work demonstrates a pathway toward accelerated AI development and deployment in healthcare.

Introduction

Healthcare in the U.S. is the world's most expensive, and the quality and safety of care do not compare well to other developed countries[1]. While electronic healthcare records are now ubiquitous in the U.S., and decision-support technologies are widely implemented, most are rule-based, and their effectiveness so far has been limited[2]. Artificial intelligence has emerged as a technique with great potential for improving care, but most organizations are not using it to any major degree. Two major limiting factors have been (1) the lack of large, labeled datasets, which are expensive and time-consuming to develop; and (2) limited system capacity to deliver recommendations to the correct clinician at the optimal time. In this manuscript, we describe a novel method called the Enhanced Transformer for Health Outcome Simulation (ETHOS), which we believe can help address many of the limitations that have prevented widespread AI adoption.

ETHOS is a novel application of the transformer deep-learning architecture, originally conceptualized for natural language processing[3]. This architecture, a cornerstone in large language model (LLM) development, is repurposed in ETHOS to analyze health-related data, moving beyond the textual focus of traditional LLMs. ETHOS is designed to process Patient Health Timelines (PHTs)—detailed tokenized chronological records of health-related events—to predict future health timelines. In PHTs, a token serves as the fundamental unit of information, encapsulating diverse data types such as patient admissions, administered medications, or time intervals. We elaborate on this pivotal aspect of our methodology in the Methods section. Our model takes the patient's health history, as represented by PHT, and subsequently forecasts future PHT (fPHT) on a token-by-token basis (refer to Fig. 1).

ETHOS's generative capabilities are gained in unsupervised learning. Once trained, ETHOS can forecast future health events without requiring task-specific training. This is done through a

zero-shot learning approach, making ETHOS a versatile foundation model for numerous healthcare applications. With appropriate modifications, ETHOS can be adapted to a broad range of data types, including but not limited to medical images, clinical and discharge notes, monitoring data, data from wearables, or omics data.

In this research, we leverage the recently released MIMIC-IV v.2.2 dataset[4,5], a rich open-source repository accompanied by our code, allowing others to replicate our findings. MIMIC-IV is expansive, chronicling more than 400,000 hospitalizations in more than 200,000 patients. Although relatively large, we anticipate that the performance of our system will further improve as we expand the dataset with additional patient histories and data types.

Importantly, we utilize the MIMIC-IV dataset in its original noisy form without any data modifications, cleaning, or targeted imputation for missing entries. The information is retained in the face of large data inconsistencies, such as discharge dates noted before admission dates. We operated under the assumption that, within large enough datasets and appropriate tokenization and training methods, ETHOS would be robust enough to handle the noisy input and automatically manage the noise/anomalies in the input data. The resilience of ETHOS to data inaccuracies and missing information has important implications for the efficiency of downstream model development. Healthcare data inevitably contains errors, some of which may not be immediately apparent or easily rectifiable. Attempts to clean large datasets can be impractical and may inadvertently introduce biases and errors. Our approach highlights the vital need for algorithms adept at managing these challenges, a prerequisite for the large-scale development of reliable and robust healthcare AI applications.

Our research showcases the zero-shot learning capabilities of ETHOS in predicting inpatient and ICU mortality, estimating ICU length of stay (LOS), and determining readmission probabilities. Additionally, we illustrate the model's versatility by performing a regression task to

estimate the first-day Sequential Organ Failure Assessment (SOFA) score[4,6] at the time of ICU admission using information before admission (see example in Figure 1). The SOFA score is a critical tool for monitoring a patient's condition in the ICU, evaluating organ function or failure across six systems—respiratory, cardiovascular, hepatic, coagulation, renal, and neurological—with each system scored from 0 to 4, culminating in a total possible minimum score of 0 and maximum score of 24. Furthermore, we predict Diagnostic-Related Group (DRG) classifications, encompassing over 771 categories, at the time of hospital discharge. The DRG system categorizes hospital cases into standardized case complexity-based Medicare and Medicaid payment groups, encouraging efficient patient care without compromising quality. The diversity of tasks ETHOS can perform, from mortality predictions and LOS estimation to SOFA scoring and DRG classification, highlights its broad applicability and zero-shot learning efficiency.

ETHOS is a foundation model[7], introducing a novel approach in the landscape of data analysis within the healthcare domain. The other foundational models developed recently have fallen into two broad categories. The first of these categories encompasses Clinical Language Models (CLaMs), a specialized subset of large language models (LLMs)[8] tailored for processing clinical and biomedical text data. These models are typically trained on extensive datasets containing clinical notes, biomedical literature, and other healthcare-related text sources. CLaMs are proficient in various clinical tasks such as extracting drug names, summarizing medical dialogues, predicting clinical outcomes, and responding to patient queries[9–13]. The second category comprises Foundation Models for Electronic Medical Records (FEMRs), representing another class of clinical foundation models tailored specifically for EMR data analysis. FEMRs undergo training on the extensive medical histories of patients, covering both structured data (such as demographics and lab results) and unstructured data (including progress notes and radiology reports). Unlike CLaMs, FEMRs are not designed to generate clinical text. Instead, they produce machine-understandable representations of patient data, facilitating tasks such as

patient phenotyping and outcome prediction[12,14–16]. Similarly, data that chronicles human lives, akin to EMR, can also be modeled effectively in this manner ([12,14–16]).

The primary distinction between ETHOS and previously published methods lies in our approach, which eliminates the need for fine-tuning or labeled data to produce accurate inferences or predictions. We demonstrate inference across a wide array of tasks without task-specific training. Moreover, the ability of ETHOS to forecast future PHTs opens the door to a wide array of bespoke and innovative applications, facilitating its use in unique scenarios in healthcare, some of them explored in the discussion section. Unlike many studies, which often apply specific criteria for selecting data for training and testing, our methodology imposes no such limitations. This feature is crucial for considering the scalability of the ETHOS approach to data sets comprising millions or even hundreds of millions of patients.

Results

Tokenization of MIMIC data and training of ETHOS

Figure 2a summarizes some statistics of the tokenization process, including the number of tokens generated and other details. Figure 2b presents visualizations of the 768-dimensional embeddings reduced to a 2D plane using Principal Component Analysis (PCA) for quantile tokens, which encode all quantitative values in the data. The tokens are arranged from Q1 (the lowest quantile) to Q10 (the highest quantile). This suggests that the transformer model has learned a sequential relationship between the tokens that mirrors their natural order, ascertaining this order from the data during the training process. The proximity between points could reflect the model's differentiation among the quantiles. We observe that the gaps between Q4, Q5, and Q6 are narrower than those between Q9 and Q10. This may suggest that the model deems the variance between population-average values to be less substantial than that

of extremely high values. For example, the difference in clinical significance between a blood pressure reading of 110 mmHg (Q5) and one of 130 mmHg (Q6) is less pronounced than the difference between 140 mmHg (Q9) and 160 mmHg (Q10), which could account for the greater disparity in the embedding vectors of high quantiles.

The embeddings for time-interval tokens, representing the approximate durations between different tokenized events in PHT, are illustrated in Figure 2c. These embeddings display a pattern analogous to that observed for Q tokens, where ETHOS systematically arranged them according to the actual time values they represent. Remarkably, the model perceives the two shortest (5m-15m, 15m-1h), and two longest (3m-6m, 6m) intervals as relatively similar.

ETHOS inferences

In our study, we conducted zero-shot inferences for a diverse array of classification tasks, including readmission to the ICU, inpatient mortality, ICU mortality, combined inpatient and ICU mortality in patients with sepsis, readmission to the ICU for patients with intracerebral hemorrhage, assignment of DRG class assessed at inpatient discharge. We also demonstrate regression of first-day SOFA score at the time of ICU admission and regression of the length of stay in ICU in days assessed upon admission. The results corresponding to these tasks are summarized in Figure 3.

To situate our results within the broader scientific discourse, we conducted a literature review, concentrating on contemporary studies that utilized the MIMIC-III and MIMIC-IV datasets for similar tasks and reported their outcomes. A notable observation from our review is that many of these studies either lacked publicly available source code or implemented specific exclusion criteria for their data selection. Such practices pose challenges for directly comparing their results with our approach. Nonetheless, we posit that the numerical outcomes reported in these works provide a valuable benchmark for assessing the performance of ETHOS.

Furthermore, we conducted a direct comparative analysis of ETHOS against specialized algorithms developed in-house, with these findings detailed in the supplementary materials.

We conducted an analysis focusing on risk estimation for inpatient and ICU mortality, calculated at the respective points of patient admission to the hospital and ICU. The test set comprised 43,309 hospital admissions with a 2.0% mortality and 7,483 ICU admissions with a 7.0% mortality. The ETHOS model demonstrated robust performance, achieving an AUC of 0.912 (95% CI: 0.898-0.922) for hospital mortality and 0.927 (95% CI: 0.914-0.938) for ICU mortality. Comparatively, in the ICU mortality risk prediction domain, the highest performance identified in our literature review was an AUC of 0.918 (95% CI: 0.915-0.922) reported by Pang et al. (2022)[17] using the XGBoost model. On the lower end, Chen et al. (2023)[18] reported an AUC of 0.642 ± 0.101. Within a specific subgroup of the test set of 3,324 patients with sepsis with 10.8% mortality prevalence, ETHOS's prediction of ICU mortality exhibited an AUC of 0.889 (95% CI: 0.870-0.906), which is a better performance than obtained in a study by Pan et al. (2023)[19], which estimated ICU mortality in adult sepsis patients using SOFA and additional features, achieving an AUC of 0.762 ± 0.006. We also estimated performance for a task of ICU mortality estimation for patients remaining in ICU for at least 24 hours in which we obtained an AUC of 0.928 (95% CI: 0.916-0.939).

Furthermore, ETHOS estimated the length of stay (LoS) in the ICU with a mean absolute error (MAE) of 2.262 days (95% CI: 2.161-2.355 days). These results paralleled those of[18], who reported an MAE of 2.42 ± 0.10 days. ICU LoS prediction and mortality risk, underscoring the competitive zero-shot performance of ETHOS across multiple key healthcare metrics.

For the ICU readmission task, ETHOS' AUC of 0.807 (95% CI: 0.786-0.827) is slightly smaller than the AUC of 0.82 obtained using knowledge graph embeddings[20] and is higher than the AUC of 0.791 (95% CI, 0.782–0.800) using LSTMs based on MIMIC-III data[21]. Additionally,

we applied our method to a task characterized by a relatively low prevalence, specifically focusing on only 174 cases of patients with hemorrhage admitted to the ICU present within our test set. The prediction of readmission by ETHOS yielded an AUC of 0.667 (95% CI: 0.402-0.839), comparable to the AUC of 0.736 (95% CI: 0.668-0.801) achieved by previous studies[22] using LightGBM. For hospital readmission, ETHOS achieved an AUC of 0.749 (95% CI: 0.743-0.755), lower than the AUC of 0.791 [0.766-0.816] obtained by Tang et al. 2023[23]. It's important to recognize that although MIMIC offers a wealth of data on acute care, it might not encompass all the subtleties necessary for readmission research, including comprehensive post-discharge outcomes or data on readmissions to various hospitals. Consequently, the accuracy of results for tasks related to readmission may be limited, regardless of the method employed.

We explored the task of predicting the first-day SOFA score at the time of admission. Given that the SOFA score is a critical indicator of survival, particularly in sepsis[6,24], this prediction can serve as a valuable indirect prognostic marker of ICU patient health status. We achieved a SOFA score estimation with an MAE of 1.502 (95% CI: 1.475-1.534). To our knowledge, no prior literature predicts first-day SOFA at the time of admission.

For the DRG assignment, we observed a top-1 (out of 771 classes) accuracy rate of 84.8% (95% CI: 84.4%-85.2%) in 28,932 hospitalizations using our methodology, a significant improvement over the 52% reported by Wang et al. (2024)[13], who explored DRG estimation using LLMs from discharge notes. This marked enhancement in performance can be attributed to the comprehensive nature of ETHOS, which incorporates a wide array of clinical events leading up to discharge within the PHT. In contrast, the approach taken by Wang et al. (2024)[13] relies solely on discharge notes, which may not encompass the breadth of information captured by PHT, thus potentially explaining the disparity in accuracy rates.

We want to reiterate an important point: all comparisons presented in this section are made between ETHOS, trained indiscriminately on the entire test population and task-specific algorithms developed using much smaller MIMIC data subsets obtained after data curation. In addition to the results in this section, in supplementary materials, we benchmark the performance of ETHOS against XGBoost[25], recurrent neural networks, and logistic regression.

Discussion

This work introduces an innovative approach to developing a Foundational Model for medical data derived from EMRs, designed to execute zero-shot inferences across a diverse range of tasks. Our model generates interpretable, causally forecasted future patient health timelines. We applied and evaluated this model using the MIMIC-IV EMR datasets, comparing its performance with the results of methods published in the literature for the same tasks. Our objective, however, was not merely to surpass the performance of these specialized SOTA implementations. Instead, we aimed to demonstrate that ETHOS, a single foundation model trained just once with zero-shot derived inference, can achieve performance levels comparable to that of multiple models optimized for various tasks. This underscores the potential of ETHOS to streamline the application of AI in healthcare by leveraging a single unified model development architecture and set of methods for multiple prediction tasks, thereby greatly enhancing medical data model development efficiency and scalability.

The application of patient timelines for generating insights has been established in existing research[12,14–16,26], as has the implementation of foundational models[7]. Our methodology sets itself apart by integrating a zero-shot capability, obviating the need for additional training beyond the initial model. Moreover, ETHOS is specifically engineered to produce causal predictions in the form of future timelines, ensuring they are inherently comprehensible to human users. This is achieved through a novel tokenization process for medical data, a distinctive feature of our work.

Another highly distinctive capability of ETHOS is the potential to generate individualized care-integrated PHT-based projected healthcare expenditures. This capability is exemplified through the prediction of Diagnosis-Related Group (DRG) codes but is not limited to this application. Specifically, ETHOS can model future PHTs at critical decision-making junctures in

patient care. For instance, ETHOS can model outcomes for administering either drug A or B, considering the patient's unique conditions (such as sex, age, race, gender, income, etc.) to determine which path might yield better clinical and cost outcomes. In this regard, ETHOS has the potential to revolutionize medical decision-analytic modeling science by incorporating a level of personalization previously unavailable in conventional decision-analytic models. This has the potential to enhance clinical decision-making and incorporate individualized real-time quantitatively robust value-based care policies into clinical care. This is a potentially transformative change, radically unlike current evidence-based medicine practices, which rely on high-quality data obtained from and averaged across patient populations[10,27,28]

In designing ETHOS, we have considered explainability, fairness, and transparency. These are vital aspects of our ongoing research. In future work, we plan to implement and test advanced visualization attention layers of the transformer[29] to gain insights into the model's reasoning process. Additionally, a dedicated interface for decision-making is envisaged further to enhance the usability of ETHOS in clinical settings.

Envisioning the development of a robust AI method that offers fully personalized advice on a wide range of medical questions necessitates learning from an extensive dataset of patients. Such a model must assimilate as much data as possible and be adaptable to a vast array of medical tasks. ETHOS represents a significant stride in this direction. Built on a transformer architecture, it is inherently scalable and, as a zero-shot learner, is versatile enough to address numerous key medical prediction tasks without task-specific training. Currently, ETHOS does not incorporate various types of critical information, including clinical and discharge notes, medical imaging and pathology images, genetic data, socioeconomic factors, lifestyle considerations, and monitoring signals. Nonetheless, the conceptual framework for incorporating these diverse data types is relatively straightforward. This can be done by leveraging the encoder and cross-attention mechanisms inherent in the transformer

architecture; we anticipate the potential for integrating a nearly limitless amount of information during training. This expansion of ETHOS's capabilities forms the cornerstone of our future work, promising to enhance its applicability and efficacy in personalized medical advice and diagnostics.

We aim to modify further and train ETHOS to apply it across diverse data sources. This capability is currently hindered by variations in data collection methodologies, disparities in data quality, and the presence or absence of certain data types across different sources. Additionally, non-overlapping populations present significant challenges, rendering ETHOS not yet generalizable. To mitigate some of these compatibility issues, we propose the development of a universal tokenization format. While this approach may resolve certain discrepancies, it does not address all underlying compatibility concerns. The ultimate solution, we believe, lies in a system capable of transforming tokenized data from one healthcare system to another, akin to text translation between languages. Specifically, for ETHOS, this would mean converting the patient journey, as encapsulated by the Patient Health Timeline (PHT), from one system's format to another. This conversion would not only facilitate a consistent and unified representation of patient histories across different systems but also offer insights into the operational nuances of these systems. Pursuing such a translation strategy represents a vital direction for our future research endeavors, alongside evaluating the methodologies introduced in this paper through analysis of prospectively collected data.

This work has limitations. We utilized the MIMIC dataset, which may be cleaner than many routine clinical datasets. Performance and usability should be tested prospectively in diverse datasets and in real-time. The transformer model in the current version of ETHOS is relatively simple and uses only 2048 PHT tokens for predictions. When token density per time is large, this may not contain sufficient information for optimal performance. Mitigation of the limitation is expected with additional computational infrastructure.

In conclusion, ETHOS presents a promising approach to deriving insights from massive clinical datasets without labor-intensive labeling or distinct model creation for each prediction task. This approach has the potential to significantly lower the costs and complexities associated with AI model development, thereby accelerating the development and implementation of healthcare AI.

Methods

Data

In this study, the Medical Information Mart for Intensive Care (MIMIC-IV) database served as a data source, providing a rich and comprehensive collection of de-identified health-related information[4]. Managed collaboratively by the Massachusetts Institute of Technology (MIT), Beth Israel Deaconess Medical Center (BIDMC), and Philips Healthcare, MIMIC-IV encompasses detailed records for more than 200,000 patients who were admitted to hospital and critical care units at BIDMC in Boston, Massachusetts, between 2008 and 2019. The following tables from the MIMIC-IV were used: 1) *Patients,* which contains static information about the patients, such as gender, date of birth, and date of death; 2) *Admissions,* which holds information about patient admissions to the hospital, including admission and discharge times, as well as information related to the hospital stay; 3) *Icustays,* which is specifically related to intensive care unit (ICU) stays, including the timings and type of ICU; 4) *Labevents,* which contains laboratory test results for patients. We used the 200 most frequent tests covering 95% of tests completed; 5) *Prescriptions,* which holds information on medications prescribed to patients during their stay, with each drug converted to ATC code[30] We converted GSN codes in MIMIC-IV to ATC codes using conversion tables[26]; 6) *Procedures* which contains information about procedures performed on patients, coded using ICD10-PCS codes; 7) *Diagnoses* which contains diagnostic information, typically coded using ICD10-CM codes. We converted ICD9 to ICD10-CM if needed using conversion table[31]; 8) *Emar,* which holds information related to the documentation and administration of medications to patients; 9) *Omr* with information about measurements taken from a patient, such as blood pressure or BMI; 10) *Services* with information about the clinical service under which a patient is managed during their hospital stay; 11) *drgcodes* DRG codes which are a classification system used in the healthcare industry to categorize hospital cases into groups that are expected to have similar hospital resource use; 12) SOFA, taken from the

derived tables in MIMIC. The remaining tables were not used in the current ETHOS implementation as they will require additional processing. For example, clinical notes require natural language processing to be converted to meaningful tokenized information.

Patient health timelines (PHTs), tokenization

The core concept behind ETHOS is the Patient Health Timeline (PHT), as depicted in Figure 1. The fundamental component of the PHT is the token, which represents a distinct unit of information occurring within the patient's health timeline. To construct the PHT, we gathered all pertinent data from tables 1 to 12 of the MIMIC-IV database, as detailed in the *Data* section. We arranged this data chronologically based on timestamps, as shown in Figure 5a, into a chronological sequence of health-related events for each patient. These events were timestamped with a floating-point number in 64-bit precision to denote the patient's age at the time of occurrence of the event. Subsequently, events from the MIMIC-IV tables were converted into tokens. Each event was represented by 1 to 7 tokens to encapsulate information about the event, as illustrated in Figure 5b. We crafted this encoding process to ensure each token conveys specific, meaningful information, with examples in Figures 5d-k. A comprehensive list of token encodings within the PHT is available in the supplementary material. The final step of tokenization involved the insertion of time-interval tokens to represent the intervals between events, depicted in Figure 5c. We employed 13 different time-interval tokens to represent the intervals. No interval token was inserted if the duration between tokens was less than 5 minutes. Typically, a single time-interval token was placed between other types of tokens unless the interval exceeded one year. In such cases, multiple 6-month tokens were used to approximate the actual interval. For example, an interval of 1.4 years was represented by three 6-month tokens, while four 6-month tokens represented 1.76 years. One interval-tokens were inserted the exact time of events was dropped from PHTs.

The patient's age and the commencement date of the PHT were represented using the same token set. We used 20 distinct tokens to denote age intervals such as 0-5 years, 5-10 years, and so forth. For instance, to encode information about a 46-year-old patient with PHT beginning in 1982, we inserted a "45_50 years" token at the 4th position in the PHT. To signify the year 1982, we used a "15_20 years" token at the 5th position of the PHT, considering 1970 as the baseline year. We emphasize that age and the commencement of the PHT are encoded in five-year intervals, given that health status typically does not undergo rapid changes with age, making finer granularity unnecessary. However, we plan to scrutinize these assumptions in subsequent research. The token denoting the commencement of the PHT delineates the temporal context of the medical data—identifying whether it corresponds to earlier medical practices (e.g., 1990s), contemporary practices, or periods in between. Using tokens with a precision of five years is done under the premise that technological and methodological progress within the medical field does not advance at a pace that justifies the necessity for time intervals more granular than five-year spans. Pertinent to the MIMIC dataset, the obfuscation of actual dates through uniform random adjustments for each patient—a measure implemented to safeguard privacy—compromises the utility of this temporal information for ETHOS, as it obscures the precise date of the start of PHT. However, the absence of precise reference dates is less critical, given that the entire dataset was collected over a relatively brief period, from 2008 to 2019[5].

As mentioned previously, token locations within the timeline are contingent upon the temporal occurrence of events. Nonetheless, certain data elements are temporally invariant, or at least presented as such within the MIMIC-IV database. In our implementation, we designate six static tokens to encapsulate the information encoded in these static data elements. Although, in reality, some of these variables may change over time, they are represented as invariable constants in the MIMIC database. We encoded this information in the six static tokens

exactly as recorded in the MIMIC dataset. These include gender, marital status, race, body mass index (BMI), birth date, and the start date of the timeline. While PHTs have the potential to extend to hundreds of thousands of tokens, our current methodology utilizes a maximum of 2048 subsequent tokens within the transformer model context, as elaborated in the "Methods: ETHOS Training" section. To accommodate invariant data, we substitute the initial six tokens of the 2048-token context with static information tokens, where the sixth token demarcates the temporal juncture of the seventh token, which is the first token of the actual timeline. Although the transformer architecture inherently facilitates the inclusion of static data via its encoder component and cross attention module[3], we opted for a more streamlined approach as described, deferring the integration of an encoder implementation to future endeavors where more substantial time-invariant data like genetics is used.

Medical encounters yield a plethora of numerical data. We employ a quantile-based tokenization strategy to process continuous numerical values, such as blood pressure readings or cholesterol levels. Specifically, all numerical values are transformed into integers representing the quantile to which each value corresponds. Quantile ranges were determined using the training dataset, where histograms of all numerical values were generated and subsequently divided into quantiles. We chose to utilize ten quantiles, a decision aimed at striking a balance between the need for precise representation of numerical data and the clinical reality that significant changes in health indicators often manifest as relatively large variations, such as shifts of 10 or 20 percent. This rationale underpins our selection of ten quantiles for tokenization.

In our study, Diagnosis-Related Group (DRG) codes for each inpatient stay were utilized, despite the absence of assigned times when they were created in the MIMIC tables. Given that a DRG code is assigned after or during discharge, we positioned it after a trio of tokens representing discharge-related information: the discharge token, a quantile token indicating the

length of the hospital stay, and a token specifying the discharge destination (e.g., home). Additionally, we incorporated data from MIMIC regarding the initial SOFA score for ICU patients, placing this token after the patient's admission-to-the-ICU token, along with a token denoting the ICU type. Given that the SOFA score in the dataset ranges from 0 to 23 (with the score of 24 never appearing), we uniformly map scores from 0-23 across 1-10 quantiles. Consequently, in quantile Q1, SOFA scores of 0, 1, and 2 (average of 1) are included, while quantile Q2 encompasses SOFA scores of 3 and 4 (average of 3.5), and this pattern continues accordingly.

ETHOS operates as a causal network. It relies solely on information available up to the time being considered in making predictions. Consequently, to ensure causality, actual values of DRG codes and SOFA scores are not employed during inference; instead, predictions of these values are used. This principle ensures that future-obtained information does not influence the prediction of yet-to-occur events. In essence, if tokens are integrated into the timeline based on their approximate occurrence time, their actual values must not be utilized for inference purposes, or they are placed in the timeline far in the future to ensure they are inserted after they occurred.

For the tokenization of drugs, whether administered or prescribed, we utilized the ATC classification system due to its hierarchical, tree-like structure. Each ATC code, comprising up to seven characters, was encoded using up to three sequential tokens: the first token for the initial three characters, the second for the subsequent character, and the third optional token, for the remaining suffix. Similarly, ICD-10-CM codes were encoded with three tokens: the first representing the first three characters of the code, the next two by the second token, and the final token capturing the code's remaining suffix. For ICD-10-PCS codes, each character in the seven-character code was represented by a distinct token. The rationale behind such tokenization is that the initial characters in those coding schemes denote specific classes of drugs and diseases or procedures, which are interpretable and have distinct meanings which

we anticipated to be important for the network's self-attention mechanisms. Looking ahead, our approach, which assigns well-defined meanings to each token, will be crucial for refining attention mechanisms and enhancing the model's explainability. This method ensures that individual tokens contribute significantly to the interpretability of the network's outcomes. For more information on the tokenization process applied to MIMIC data in our analysis, as well as examples of Patient Health Timelines (PHTs), readers are directed to Figure 6 and the supplementary materials where we present real PHTs used in this work with annotations.

ETHOS training

We employ a model inspired by the decoder architecture of the transformer[3], drawing parallels between tokenized text in Natural Language Processing (NLP) and our approach to tokenizing PHTs. The ETHOS model's training begins by synthesizing a dataset from existing patient records. Each patient's PHT is ended with a "End of timeline" token, and then they are concatenated, creating a single long sequence of tokens for the training. Similarly to generative LLM, ETHOS is trained to predict a single token based on the context of preceding ones. Given the large data scale and model complexity, this phase is resource-intensive similar to methods for training used for NLP transformers used in LLMs[3,32]. We estimated that the size of the network training task that we face with ETHOS is similar to GPT-2[8], and therefore we used the size of the transformer used in that network as a starting point. We made heuristic adjustments to the size of the network to optimize the value of the loss function. Further details on our training methodology of transformers are provided in Brown et al. (2020)[8] and for our implementation in supplementary material.

ETHOS inference

During inference, ETHOS functions analogous to a document completion tool in which word sequences instead of health-related events are sequenced into a PHT. The procedure begins with the patient's history recorded in their PHTs. The last 2048 tokens—or the entire PHT if it contains fewer than 2048 tokens—are used to initiate the inference in the current ETHOS implementation. ETHOS then generates one token at a time through the following steps: (1) generating an array of probabilities for all potential tokens, (2) stochastically selecting a new token based on these probabilities, (3) appending the new token to the sequence while removing the oldest one to maintain the context size at 2048 tokens, (4) go to 1. This generative sequence proceeds until it encounters predefined stopping conditions, which may include the appearance of a token showing the patient's death or the sum of time-interval tokens surpassing a certain threshold. Additional stopping criteria may be established. The stochastic nature of this method allows for the creation of multiple future PHTs (fPHTs). Multiple fPHTs are used to assess uncertainties as each of the fPHTs represents an alternative prediction of the future.

Evaluation of Clinical Outcomes and Tasks Using ETHOS

The experiments were chosen so the results can be compared to the work of others in terms of the estimation of inpatient mortality and readmission on MIMIC data. Patients in the MIMIC were randomly divided into training and testing groups, with splits of 90%/10%.

The chance of inpatient mortality was assessed at the time of admission for all inpatient stays for patients in the test set unless the discharge day was unknown. This was performed by the generative process that began with the admission token and ended upon generating a discharge or death token, repeating this cycle 20 times. The 'N', representing the number of times a death token was generated first, was divided by 20 to estimate the chance of inpatient mortality. Similarly, the likelihood of ICU mortality was computed for the MIMIC dataset, with an additional experiment conducted where predictions were made starting 24 hours after ICU

admission, rather than at the point of ICU admission. In the same simulation, the LOS in the ICU was estimated by aggregating the time-interval tokens generated in the simulated timeline until the discharge token appeared. Instances where the patient died in the ICU during the simulation were excluded from the LOS calculation. We opted for 20 repetitions, yielding 21 unique probability estimators, which were adequate for constructing robust Receiver Operating Characteristic (ROC) curves yielding excellent Gaussian fits (Figure 3). Nevertheless, alternative repetition counts may also be employed.

To calculate the probability of 30-day inpatient readmission, the generation of fPHTs commenced at the discharge token from inpatient stays and ceased upon the appearance of either a new admission or death token or when the cumulative time tokens generated exceeded 30 days. The simulation was repeated 20 times. The probability of 30-day readmission was then derived as M/20, where 'M' is the count of terminations occurring because of patient new admission tokens across the 20 repetitions.

In our approach, tasks are accomplished by simulating future patient health timelines. Yet, ETHOS offers additional methods for deriving insights, two of which we illustrate here. For instance, in the construction of PHTs following each ICU admission, a sequence is created starting with a token that identifies the type of ICU, followed by a SOFA score token, and then by a Q token that signifies the actual SOFA score on the first day. We predict the SOFA score using SOFA Q node probabilities as generated by ETHOS and the mean SOFA score per quantile as assigned during tokenization (Figure 4a).

The exact timing of the 1-day SOFA score assessment is not specified in the dataset, leading to a potential causality issue by inserting the SOFA score immediately after admission, as it relies on data acquired subsequently. During the model's training phase, ETHOS permits this apparent causality violation. However, such true values of 1-day SOFA scores, not available at

the moment of ICU admission, are not used for simulating future timelines during inference to prevent causality violation during inference. Instead, these scores are predicted from prior information, as demonstrated in our study. This feature of ETHOS enables the inclusion of information with indeterminate timing.

Another distinctive inference capability facilitated by ETHOS is DRG class estimation. As illustrated in Fig. 4b, the token denoting the DRG class is consistently positioned following the discharge token and a Q token specifying the length of hospital stay. With 771 unique tokens available for this purpose, we infer the actual class by generating a probability array in the final network layer of the transformer for the DRG token. This array is then utilized to predict the classification's top-1 and top-2 accuracy metrics.

Statistical Analysis

The performance of classification algorithms of binary tasks was assessed using Receiver Operating Curve Analysis (ROC). The ROC curves were fitted to experimental points using Gaussian models with unequal variances for binary hypotheses (code provided). Values of Areas Under Curves (AUCs) and 95% confidence intervals (CI) were calculated using bootstrapping (code provided). For multiclass classification (DRG task), we used top-1 and top-2 accuracy. We used mean absolute error (MEA) for the regression tasks to indicate prediction fidelity with 95% confidence intervals estimated using bootstrapping. Python numpy and scikit-learn were used.

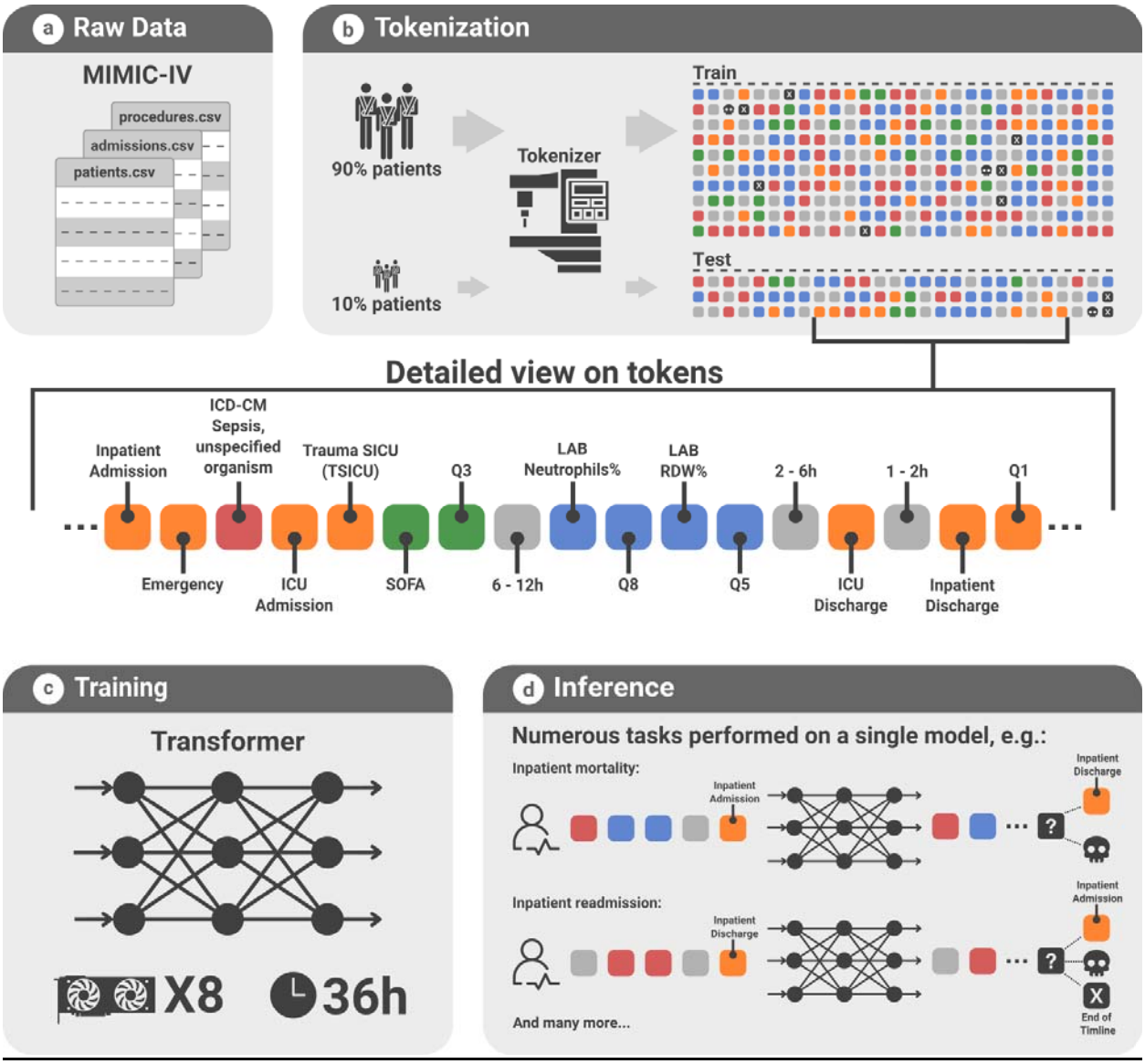


Figure 1: Implementing the ETHOS Model with EMR Data.

a) Extraction of raw patient data from the MIMIC-IV database, encompassing tables of admissions, patient demographics, medical procedures, among others.

b) The tokenization process, utilizing data from 90% of patients for model training and the remaining 10% for testing, transforms complex medical records into structured PHT for efficient model processing.

c) Training phase illustration, employing a transformer architecture optimized across 8 GPUs over a span of 36 hours.

d) Demonstration of ETHOS's zero-shot inference capabilities, highlighting its proficiency in performing tasks such as predicting inpatient mortality and readmission rates, leveraging forecasted future PHTs.

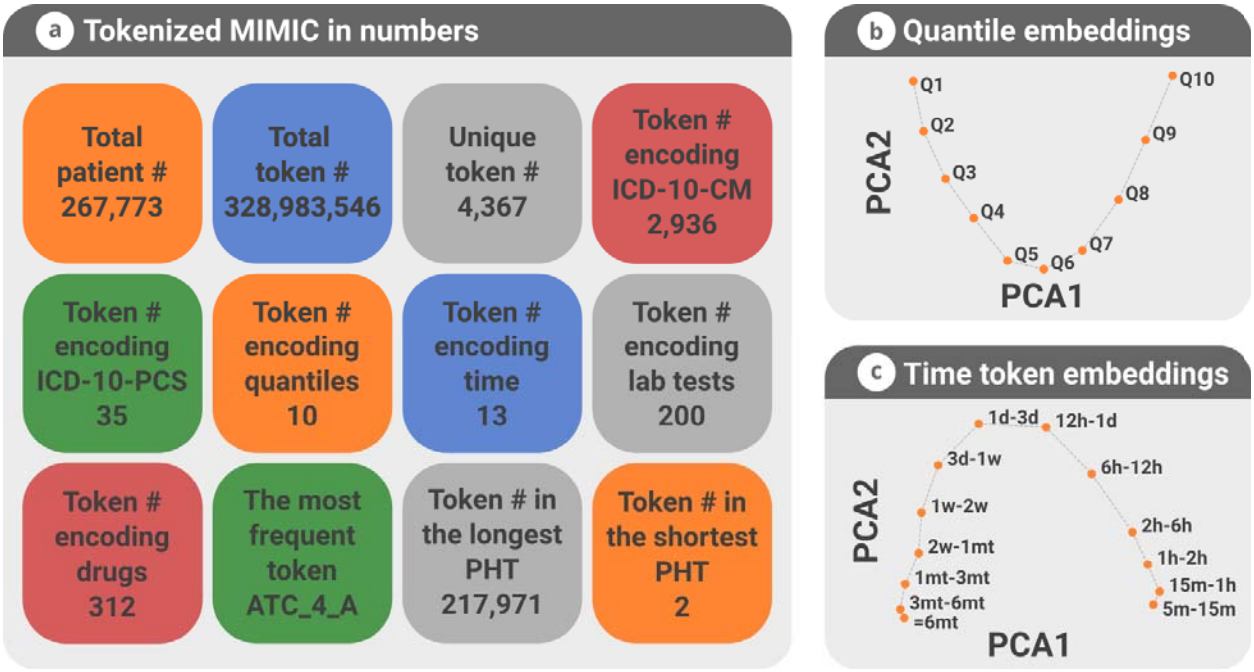


Figure 2. Tokenization and Embedding Visualizations of MIMIC-IV Data.

a) Overview of key insights derived from the tokenization process applied to MIMIC-IV data.

b) Visualization of embedding vectors for quantile tokens (Qs), which categorize quantitative information across the dataset. Each quantitative measure (e.g., blood pressure) is encoded by a preceding category-specific token followed by a quantile token, delineating its position within a predefined value range. This method facilitates a structured, scalable representation of complex data types via a systematic token sequence.

c) Visualization of embedding vectors for time-interval tokens, illustrating the temporal distribution and relationships within the PHT.

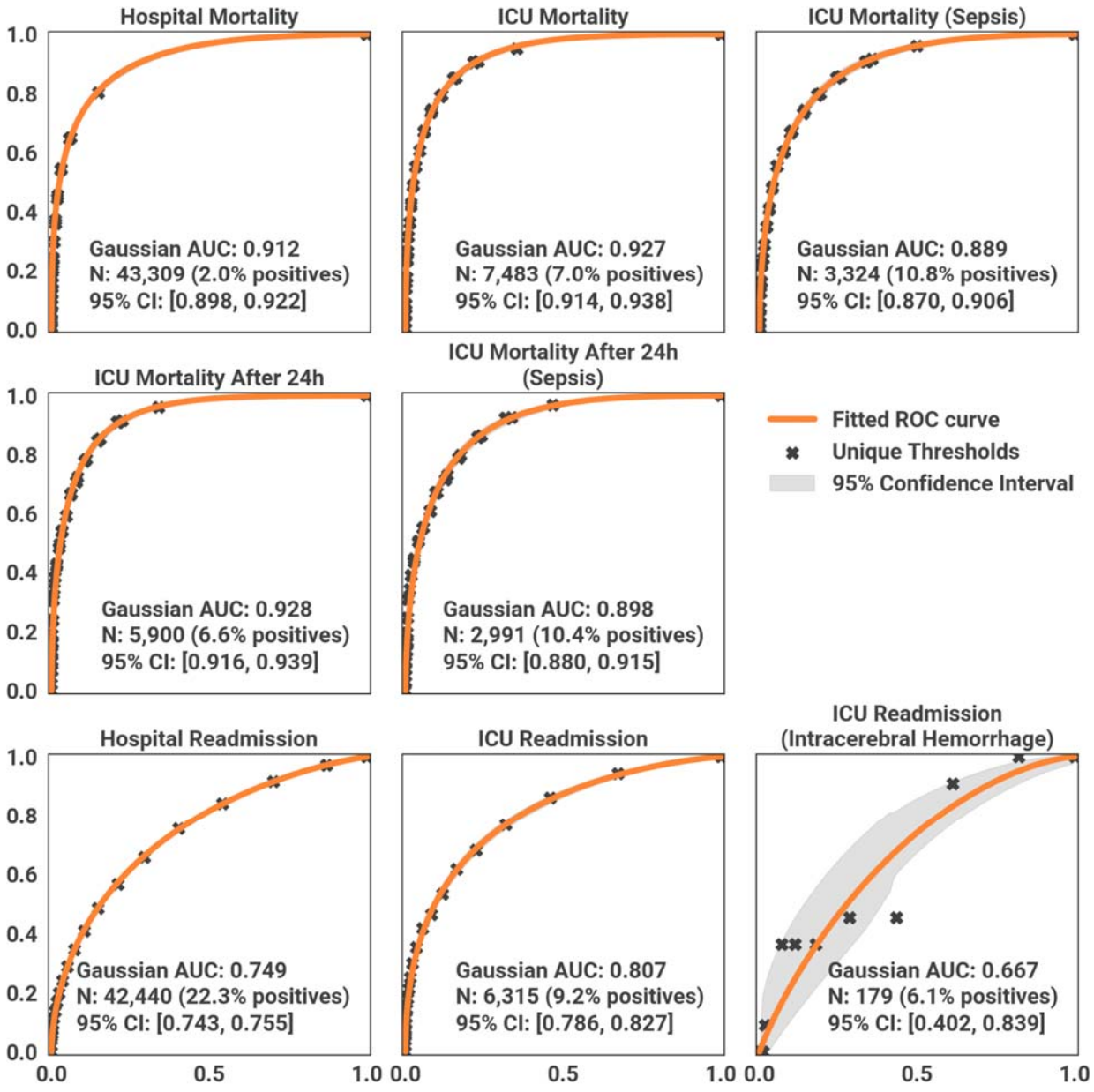


Figure 3: Receiver Operating Characteristic (ROC) Curves for Predictive Tasks via the ETHOS Model.

Each graph delineates the model's efficacy in forecasting distinct clinical outcomes, specifically mortality and readmission rates. Accompanying each ROC curve are the case count (N), the outcome prevalence, and the 95% confidence interval for the AUC. Points marked with an 'X' denote specific thresholds utilized for classification decisions within the ETHOS model.

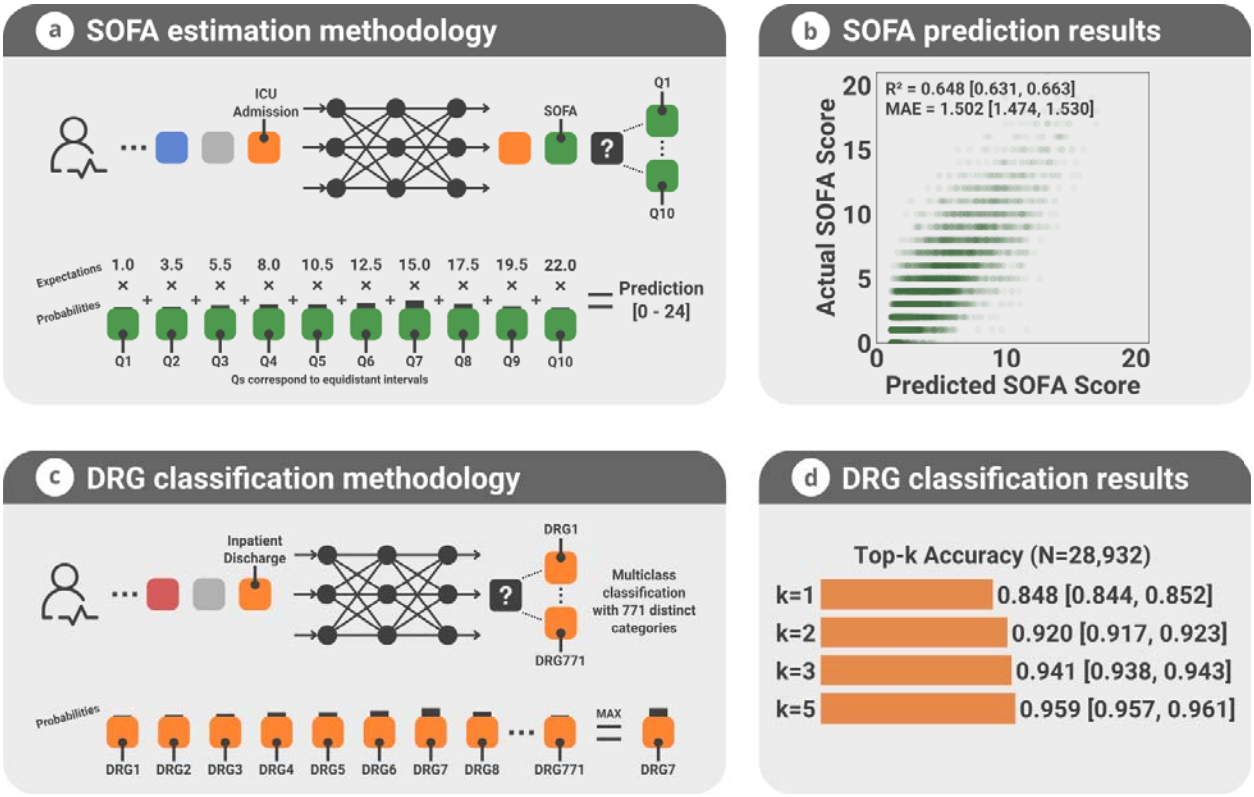


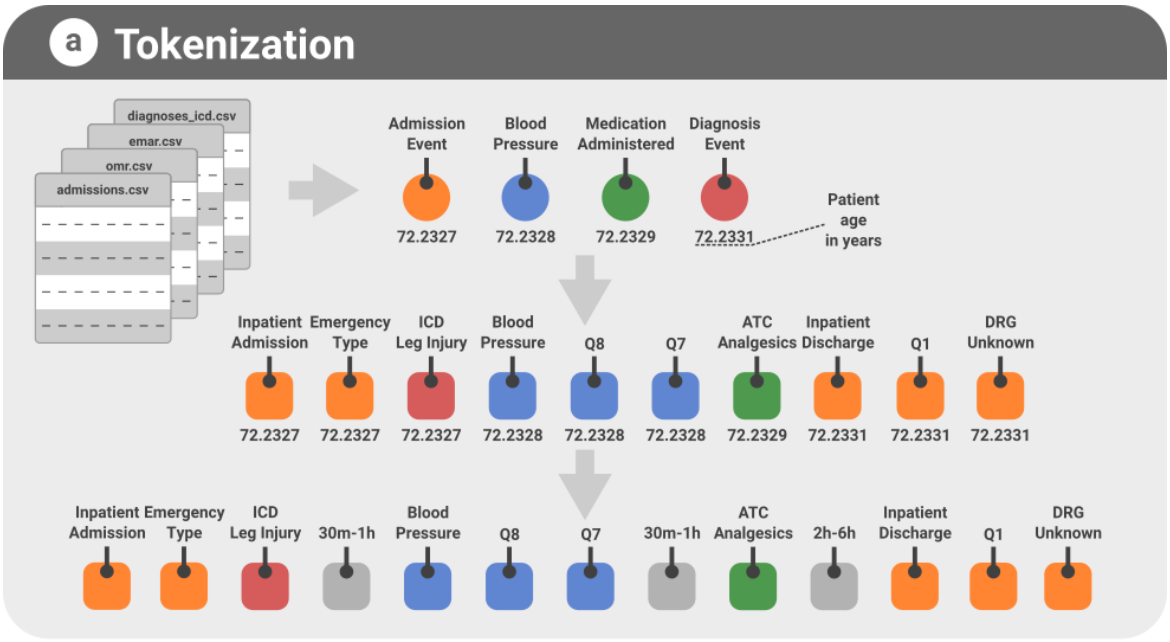
Figure 4: ETHOS Model Performance on SOFA Estimation and DRG Classification.

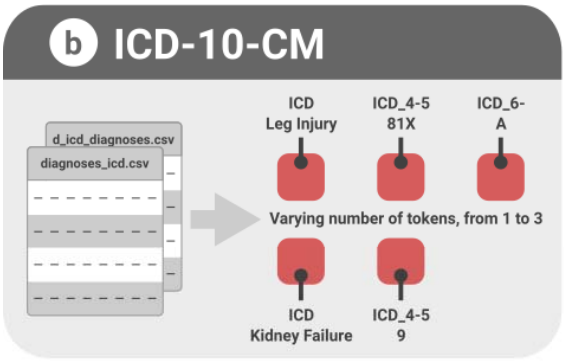
a) Estimation of the first-day Sequential Organ Failure Assessment (SOFA) score at ICU admission by ETHOS, which generates a sequence of three tokens: the admission type (orange token), a SOFA token (indicating the SOFA score estimation will follow), and a quantile token (q-token indicated by question mark) predicting probabilities of the SOFA score's quantile, as detailed at the bottom of the panel (a). The fixed position of the SOFA token ensures its consistent prediction immediately after ICU admission. The SOFA score is derived using quantile probabilities generated by ETHOS and average value of SOFA for ten quantiles (values of 1.0, 3.5 …). Since SOFA value 24 was not present in the dataset we predict values 0-23.

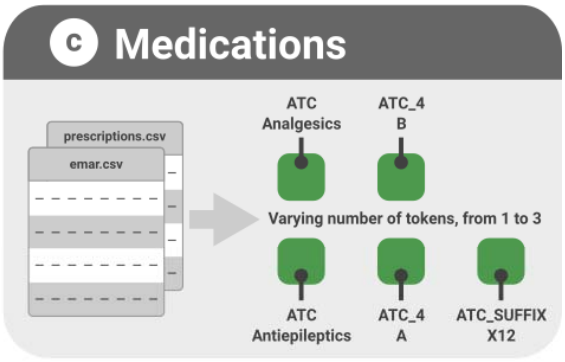
b) Correlation plot between actual and predicted SOFA scores.

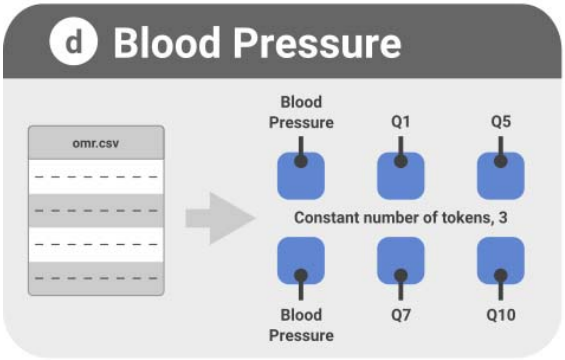
c) For Diagnostic Related Groups (DRG) classification. The model is trained to insert a DRG token after tokens typically used at discharge time, utilizing a placeholder “DRG_UNKNOWN” for if DRG is unknown in the training set. Predicted probabilities are used to compute the top-{1,2,3,5} DRG classifications.

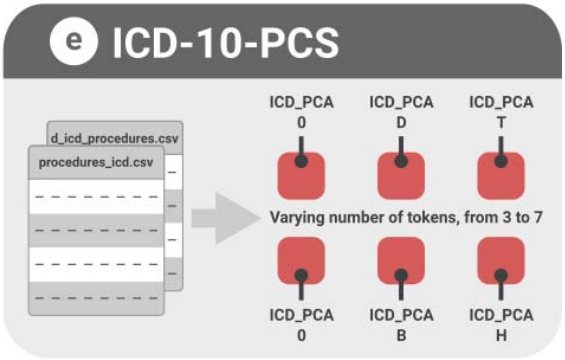
d) Visualization of DRG classification accuracy, showcasing the model's predictive performance.

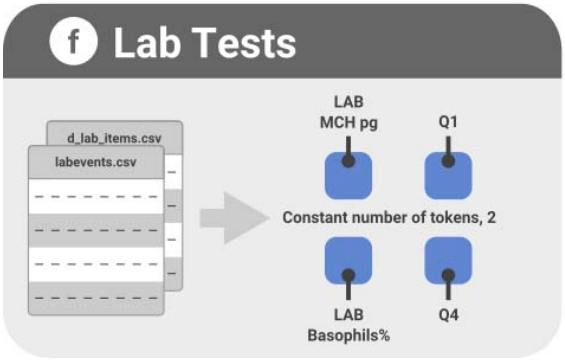












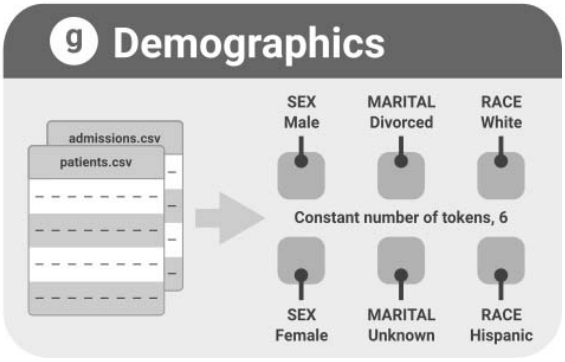


Figure 5: Stages of PHT Construction and Tokenization in ETHOS

Figure 5. a) The process begins with assembling a chronological list of events from MIMIC-IV tables, Each entry on the list is time stamped with 64-bit real value only 6 significant digits show for clarity, indicating the patient's age at which the event occurred. b) Subsequently, list elements are transformed into tokens using ETHOS tokenization scheme. Based on the event's nature, one event can be translated into 1 up to 7 tokens. Each token derived from the same event shares its timestamp. c) The final step involves representing time gaps between events by inserting time-interval tokens. If the time difference between events is less than 5 minutes—the minimum value represented by the token for the shortest time interval—no token is added. After adding interval-tokens, timestamps are stripped from the timeline. Not pictured is insertion of 6 static-information tokens at the beginning of timeline which is the last step in construction of PHT. Panels d-k showcase various examples of information encoding via tokens: (d) Depending on the ICD-10-CM code, 1 to 3 tokens are utilized for representation, with the first token corresponding to the code's first three characters, the fourth and fifth characters possibly represented by another token, and an optional third token for the remaining characters in the ICD code. (e) Medications, coded by ATC codes, are similarly encoded by 1 to 3 tokens based on the specificity of the code, with the first token representing the first three characters, the second for the next two, and the third for the remaining characters. (f) Blood pressure measurements are consistently encoded using three tokens: one to indicate the BP measurement and two quantile tokens for systolic and diastolic pressure values, respectively. (g) ICD-PCS codes may be represented by up to seven tokens, with each token denoting one character of the code. (h) Lab tests are depicted by a token that describes the type of test followed by a quantile token for the test's numerical value. Finally, (i) demographics are depicted which are part of static tokens, always positioned at the beginning of the PHT.

Patient Health Timeline (PHT)	Notes
ED_ADMISSION_START----------	Admission to Emergency Department
_2h-6h----------	Time interval that elapsed after admission to ED was recorded and next event
LAB_pH_units----------	Results Lab test of pH and the unit of this result is units. The same lab test may have different tokens due to units
_Q4----------	Quantile token referring to the result of lab test of pH
LAB_Protein_mg/dL	The next tokens are plethora to tokens indicating various lab results
_Q4	
LAB_Absolute Monocyte Count_K/uL	
_Q6	
_6h-12h----------	Token indicating that another 6-12 hour period elapsed
INPATIENT_ADMISSION_START----------	Token indication an admission to the hospital
TYPE_OBSERVATION----------	The type of the admission is observation - this token always follows admission token
INSURANCE_MEDICARE----------	The patient's insurance is MEDICARE - this token always follow
ICD_Other symptoms and signs involving------	The primary diagnosis at the beginning of the hospital stay is Altered mental status, unspecified (R4182), which is
cognitive functions and awareness	broken down to two tokens; R41 and 82. This token represent the first three characters
ICD_4-5_82----------	This represents the 4th and 5th character of ICD code from the previous token.
ICD_Dorsalgia	More ICDs representing diagnoses follow.
ICD_4-5_16	
ICD_Malaise and fatigue	
ICD_4-5_1	
ICD_Personal history of certain other diseases	
ICD_4-5_73	
ICD_Other hypothyroidism	
ICD_4-5_9	
ICD_Type 2 diabetes mellitus	
ICD_4-5_21	
ICD_Dorsalgia	
ICD_4-5_5	
ICD_Abnormalities of gait and mobility----------	The last ICD in the group.
ICD_4-5_2	
TRANSFER_MED----------	Transfer to a different care unit - MED
_6h-12h----------	Time interval between 6h and 12h
ATC_stomatological preparations----------	Token indicating medication, in this case ATC code: A01AD05 broken down in PHT into 3 tokens; A01, A, D05
ATC_4_A	
ATC_SUFFIX_D05	
ATC_diuretics----------	Another medication with ATC code broken into 3 subsequent tokens
ATC_4_A	
ATC_SUFFIX_A03	
ATC_agents acting on the renin-angiotensin system--	Another medication with ATC code broken into 3 subsequent tokens
ATC_4_A	
ATC_SUFFIX_A03	
_6h-12h	
ED_ADMISSION_END----------	Discharge from Emergency Department
_Q10----------	Quantile quantifying the length of the stay in Emergency Department, always put at end of admissions, icu, ed
INPATIENT_ADMISSION_END----------	Discharge from the hospital
_Q2----------	Quantile quantifying the length of the stay in the hospital based on all stays in the data
DISCHARGED_UNKNOWN----------	The reason for discharge. In this case the reason is unknown.
UNKNOWN_DRG----------	After all admissions, a DRG-class token is inserted. Here it is unknown.
_=6mt----------	Time interval of 6 months, which indicates that no information is available for approximately 6 months
LAB_Epithelial Cells_#/hpf----------	After approximately 6 month patient underwent lab test which is indicated by this token

Figure 6: Example segments of an actual PHT

For enhanced understanding, we utilize descriptive labels for token names, although within ETHOS, these are internally represented as integers. A short abbreviated section is provided here and more extensive examples are provided in supplementary materials.

Data availability

The MIMIC-IV dataset is publicly available at https://physionet.org/content/mimiciv/2.2/. The code, ETHOS model weights used for all inferences, results of inferences, scripts to generate numerical results for all aspects of this study for the MIMIC-IV will be made available upon reasonable request after peer review and publication.

Declaration of interests

YJ is currently also affiliated with Verily life science, SSF, CA.

References

1. Schneider, E. C. *et al.* Reflecting Poorly: Health Care in the US Compared to Other High-Income Countries. *New York: The Commonwealth Fund* (2021).
2. Bates, D. W. *et al.* 'Improving smart medication management': an online expert discussion. *BMJ Health Care Inform* **29**, (2022).
3. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
4. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* **10**, 1 (2023).
5. Johnson, A. *et al.* Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/2.2/(accessed Oct 1, 2023)* (2023).
6. Raith, E. P. *et al.* Prognostic Accuracy of the SOFA Score, SIRS Criteria, and qSOFA Score for In-Hospital Mortality Among Adults With Suspected Infection Admitted to the Intensive Care Unit. *JAMA* **317**, 290–300 (2017).
7. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
8. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **abs/2005.14165**, (2020).
9. Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med* **6**, 135 (2023).
10. Zack, T. *et al.* Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health* **6**, e12–e22 (2024).
11. Li, F. *et al.* Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR Medical Informatics* **7**, e14830 (2019).
12. Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).

13. Wang, H., Gao, C., Dantona, C., Hull, B. & Sun, J. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ Digit Med* **7**, 16 (2024).
14. Steinberg, E. *et al.* Language models are an effective representation learning technique for electronic health record data. *J. Biomed. Inform.* **113**, 103637 (2021).
15. Li, Y. *et al.* Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. *IEEE J Biomed Health Inform* **27**, 1106–1117 (2023).
16. Savcisens, G. *et al.* Using sequences of life-events to predict human lives. *Nat Comput Sci* **4**, 43–56 (2024).
17. Pang, K., Li, L., Ouyang, W., Liu, X. & Tang, Y. Establishment of ICU Mortality Risk Prediction Models with Machine Learning Algorithm Using MIMIC-IV Database. *Diagnostics (Basel)* **12**, (2022).
18. Chen, J., Qi, T. D., Vu, J. & Wen, Y. A deep learning approach for inpatient length of stay and mortality prediction. *J. Biomed. Inform.* **147**, 104526 (2023).
19. Pan, X. *et al.* Evaluate prognostic accuracy of SOFA component score for mortality among adults with sepsis by machine learning method. *BMC Infect. Dis.* **23**, 76 (2023).
20. Carvalho, R. M. S., Oliveira, D. & Pesquita, C. Knowledge Graph Embeddings for ICU readmission prediction. *BMC Med. Inform. Decis. Mak.* **23**, 12 (2023).
21. Lin, Y.-W., Zhou, Y., Faghri, F., Shaw, M. J. & Campbell, R. H. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS One* **14**, e0218942 (2019).
22. Miao, J. *et al.* Predicting ICU readmission risks in intracerebral hemorrhage patients: Insights from machine learning models using MIMIC databases. *J. Neurol. Sci.* **456**, 122849 (2024).
23. Tang, S. *et al.* Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks. *IEEE J Biomed Health Inform* **PP**, (2023).

24. Minne, L., Abu-Hanna, A. & de Jonge, E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. *Crit. Care* **12**, R161 (2008).
25. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
26. Bornet, A. *et al.* Comparing neural language models for medical concept representation and patient trajectory prediction. *medRxiv* (2023).
27. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
28. Abid, A., Farooqi, M. & Zou, J. Large language models associate Muslims with violence. *Nature Machine Intelligence* **3**, 461–463 (2021).
29. Vig, J. A Multiscale Visualization of Attention in the Transformer Model. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (eds. Costa-jussà, M. R. & Alfonseca, E.) 37–42 (Association for Computational Linguistics, 2019).
30. WHO. Anatomical Therapeutic Chemical (ATC). *WHO Collaborating Centre for Drug Utilization Research* www.whocc.no.
31. ICD10 codes. *Centers for Medicare & Medicaid Services* https://www.cms.gov/medicare/coding-billing/icd-10-codes.
32. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).