AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

DZIEDZINA: NAUKI INŻYNIERYJNO-TECHNICZNE

DYSCYPLINA: Informatyka techniczna i telekomunikacja

ROZPRAWA DOKTORSKA

Model Bazowy Elektronicznej Dokumentacji Medycznej

Autor: Paweł Renc

Promotor rozprawy: Prof. dr hab. inż. Jarosław Wąs, AGH w Krakowie
Drugi promotor: Dr. Arkadiusz Sitek, Harvard Medical School

Praca wykonana: AGH w Krakowie, Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Kraków, 2025

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

FIELD OF SCIENCE: ENGINEERING AND TECHNOLOGY

SCIENTIFIC DISCIPLINE: Information and communication technology

DOCTORAL DISSERTATION

Foundation Model for Electronic Health Records

Author: Paweł Renc

First supervisor: Prof. dr hab. inż. Jarosław Wąs, AGH University of Krakow
Second supervisor: Dr. Arkadiusz Sitek, Harvard Medical School

Completed at: AGH University of Krakow, Faculty of Electrical Engineering, Automatics, Computer Science, and Biomedical Engineering

Kraków, 2025

Abstract

The central hypothesis of this doctoral dissertation is that *tokenized patient health timelines, modeled through generative transformer architectures, provide a universal representation of electronic health records (EHRs) that enables foundation models to achieve scalability, privacy preservation, and clinical fidelity*. This hypothesis arose from the conviction that healthcare artificial intelligence must move beyond narrow task-specific models toward general-purpose frameworks that unify prediction, simulation, and deployment.

To test this hypothesis, three major contributions were developed. First, the Enhanced Transformer for Health Outcome Simulation (ETHOS) introduced a new paradigm in which longitudinal health records are represented as sequences of tokens, forming patient health timelines (PHTs). ETHOS demonstrated for the first time that zero-shot generative modeling could achieve predictive performance across diverse clinical tasks—including mortality, readmission, length of stay, SOFA score estimation, and DRG classification—without requiring fine-tuning. This validated that tokenized PHTs and generative transformers can serve as a single foundation model architecture for heterogeneous tasks, thereby proving the feasibility of the first component of the hypothesis.

Second, ETHOS was extended into the Adaptive Risk Estimation System (ARES). ARES leveraged the generative capacity of ETHOS to simulate multiple possible futures for a given patient and derive adaptive, personalized, and explainable risk estimates. This work demonstrated that adaptive inference over simulated trajectories can provide real-time, clinically meaningful explanations, confirming the second component of the hypothesis.

Third, Federated Timeline Synthesis (FTS) addressed the challenge of multi-institutional model training in healthcare, where privacy concerns prevent data sharing. FTS introduced a methodology for generating synthetic timelines at local sites, which are then aggregated centrally for training. Extensive fidelity and privacy evaluations showed that models trained on federated synthetic data preserved predictive performance while safeguarding sensitive information. This validated the final component of the hypothesis, proving that privacy-preserving federated synthesis is a viable mechanism for cross-institutional deployment.

The impact of this work has already been significant. ETHOS has rapidly gained adoption, with over 30 citations within its first year of publication. It has been extended by leading organizations such as Microsoft Research and Epic Systems, who validated its scalability across hundreds of millions of patients and billions of events. The framework has also inspired interest from global technology companies and health technology startups, demonstrating its practical and scientific importance.

In conclusion, this dissertation has proven the hypothesis that tokenized patient timelines, modeled with generative transformers, provide a universal representation of EHRs capable of advancing prediction, simulation, explainability, and federated deployment. These contributions mark a decisive step toward building foundation models that are general-purpose, interpretable, privacy-preserving, and deployable at scale, laying the groundwork for the next generation of trustworthy clinical AI.

Streszczenie

Główną hipotezą tej pracy doktorskiej jest założenie, że *tokenizowane chronologie zdarzeń zdrowotnych pacjentów, modelowane przy użyciu generatywnych architektur transformatorowych, mogą stanowić uniwersalną reprezentację elektronicznej dokumentacji medycznej (EHR). Takie podejście pozwala na budowę modeli bazowych, które są skalowalne, zapewniają ochronę prywatności oraz zachowują wierność kliniczną*. Hipoteza ta wynika z przekonania, że sztuczna inteligencja w medycynie powinna wyjść poza wąsko zdefiniowane modele zadaniowe i zmierzać w stronę ogólnych ram, które łączą predykcję, symulację i możliwość wdrażania w różnych instytucjach.

Aby zweryfikować to założenie, opracowano trzy główne rozwiązania. Pierwszym z nich jest Enhanced Transformer for Health Outcome Simulation (ETHOS), który wprowadził nowy sposób reprezentowania danych klinicznych jako sekwencji tokenów tworzących chronologie zdrowotne pacjentów (PHT). ETHOS po raz pierwszy pokazał, że generatywne modelowanie w trybie zero-shot może osiągać wysoką skuteczność predykcyjną w wielu zadaniach klinicznych — takich jak przewidywanie śmiertelności, rehospitalizacji, długości hospitalizacji, oceny SOFA czy klasyfikacji DRG — bez konieczności dostrajania modelu. Tym samym potwierdzono, że tokenizowane PHT i transformatory generatywne mogą być uniwersalną architekturą dla zróżnicowanych zadań.

Drugim wkładem jest Adaptive Risk Estimation System (ARES), rozwinięcie ETHOS oparte na symulacji wielu możliwych scenariuszy przyszłości pacjenta. Dzięki temu ARES potrafi generować adaptacyjne, spersonalizowane i wyjaśnialne prognozy ryzyka w czasie rzeczywistym. Wykazano w ten sposób, że symulacja przyszłych trajektorii może służyć jako narzędzie do praktycznych i zrozumiałych klinicznie ocen ryzyka, co potwierdziło drugi element hipotezy.

Trzecim rozwiązaniem jest Federated Timeline Synthesis (FTS), które rozwiązuje problem trenowania modeli w różnych instytucjach medycznych bez konieczności udostępniania wrażliwych danych. FTS umożliwia generowanie syntetycznych chronologii pacjentów lokalnie, a następnie ich łączenie do treningu modelu globalnego. Szczegółowe testy pokazały, że takie podejście pozwala zachować wysoką skuteczność predykcyjną przy jednoczesnym zapewnieniu ochrony prywatności, co potwierdziło trzeci element hipotezy.

Znaczenie tej pracy widać również w jej szybkim oddźwięku. ETHOS w ciągu roku od publikacji został zacytowany ponad 30 razy i zyskał miano podejścia przełomowego. Został zaadaptowany przez takie instytucje jak Microsoft Research czy Epic Systems, które pokazały jego skalowalność na setkach milionów pacjentów i miliardach zdarzeń medycznych. Co więcej, rozwiązanie to wzbudziło zainteresowanie globalnych firm technologicznych (m.in. Google) oraz mniejszych firm ubezpieczeniowych i medtechowych, co potwierdza zarówno jego naukową, jak i praktyczną wartość.

Podsumowując, w pracy dowiedziono, że tokenizowane chronologie pacjentów, modelowane transformatorami generatywnymi, mogą być uniwersalną reprezentacją EHR. Umożliwia to tworzenie modeli bazowych, które są skalowalne, interpretowalne, chronią prywatność i nadają się do wdrożeń na szeroką skalę. Wyniki te stanowią ważny krok w kierunku nowej generacji wiarygodnej sztucznej inteligencji w medycynie.

Contents

Acknowledgments

This dissertation is not only the outcome of years of research but also of my personal journey. Along the way, I have been fortunate to meet people who believed in me, guided me, and shaped the path that brought me here.

I want to begin by thanking Prof. Jaroslaw Was. You were the first to nudge me into research, encouraging me to write my very first scientific article during my bachelor studies. You also gave me the opportunity to go on a student exchange to Italy—an experience that completely changed my life and set me on the path toward a PhD. I am also grateful to Prof. William Spataro and Dr. Alessio De Rango from the Università della Calabria. You planted in me the seed of research curiosity and made me realize that I truly wanted to pursue the PhD degree.

During my Master's studies, I had the privilege of working with Dr. Patryk Orzechowski at the University of Pennsylvania. Thank you, Patryk, for awakening my devotion to research, for introducing me to the world of scientific conferences, and for sharing so many stories about how research in the United States works—about grants, collaborations, and the bigger picture. You encouraged me to think broadly, to expand my network, and to grow as a young researcher.

I would also like to thank Dr. Maciej Besta for his mentorship. Maciej, you showed me science through the lens of a world-class HPC laboratory and taught me lessons that went far beyond technical skills. From you, I learned perseverance, self-esteem, and the mindset it takes to succeed. Your guidance and the exposure you gave me to international research environments were invaluable, and they remain with me to this day.

Most importantly, I want to thank Prof. Arkadiusz Sitek, my mentor and friend. Arek, you believed in me from the very beginning and gave me the opportunity to spend over two years at Massachusetts General Hospital and Harvard Medical School. Those years were priceless, filled with constant lessons in both science and life. They had an immense impact on me and shaped me into the researcher I am today. Thank you for the countless hours of discussions, even seven days a week when needed, for instilling in me positivity, for teaching me the value of scientific doubt, and for standing by me in the most difficult moments. Your support and belief in me gave me the strength to persevere through the hardships of research, and I know I would not be here without you.

Once again, I wish to thank Prof. Jaroslaw Was, whose invaluable support extended far beyond the beginning of my journey. You helped me navigate the challenges with the university and made it possible for me to pursue internships and research stays abroad. Without your understanding and guidance, I would never have been able to realize my PhD in such a unique way. I am deeply indebted to you.

On a personal note, I am deeply grateful to my family and friends. Your unwavering belief in me, even through years of distance and uncertainty, gave me the strength to keep going. This journey often unfolded far away from home, and adapting to life abroad was not easy. The cultural and institutional differences, the distance from loved ones, and the struggle of building a new life were daunting at times. Yet every challenge made me stronger and taught me resilience. Thanks to you, I became a person with mental strength like steel—-something I needed more than ever during this PhD.

Year	Title	Where published / submitted	#Authors, my pos
2025	**Paweł Renc et al. Federated Timeline Synthesis: Scalable and Private Methodology For Model Training and Deployment**	**preprint (under review)**	**6, 1**
2025	**Paweł Renc et al. Foundation Model of Electronic Medical Records for Adaptive Risk Estimation**	**GigaScience (journal) IF=3.9, TOP10 Scopus, 200 pts**	**10, 1**
2025	M. Lemoli, P. Renc et al. Proteins predicting arterial stiffness and high blood pressure and their association with cardiovascular outcomes: Results from UK Biobank	Atherosclerosis (journal) **IF=5.7, TOP10 Scopus/Wos, 140 pts**	8, 2
2025	MK Grzeszczyk, P. Renc et al. RegScore: Scoring Systems for Regression Tasks	International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) **Core A, 140 pts**	5, 2
2025	What Clinicians Want AI to Measure, and How They Want It Done	under review	8, 5
2024	Meds decentralized, extensible validation (meds-dev) benchmark: Establishing reproducibility and comparability in ml for health	Machine Learning for Health Conference (ML4H)	
2024	N. Oufattole, T. Bergamaschi, P. Renc et al. Meds-torch: An ml pipeline for inductive experiments for ehr medical foundation models	NeurIPS Workshop on Time Series in the Age of Large Models	6, 3
2024	**Paweł Renc et al. Zero shot health trajectory prediction using transformer**	**npj Digital Medicine (journal) IF=15.1 TOP10 Scopus/WoS**	**7, 1**
2023	M. Besta, P. Renc et al. High-performance and programmable attentional neural network accelerators	International Conference for High Performance Computing, Storage, .. **Core A, 140 pts**	18, 2
2023	A. Sitek, P. Renc, Q Li, Online Reconstruction of Adaptive Whole-Body SPECT	IEEE Nuclear Science Symposium and Medical Imaging Conference	3, 2
2023	EC Murray, C Delles, P Orzechowski, P Renc, et al. Vascular phenotypes in early hypertension	Journal of Human Hypertension **IF=3.4, Q1 Scopus/WoS, 100 pts**	7, 4
2022	Paweł Renc et al. Towards efficient GPGPU cellular automata modeling	Journal of Computational Science **IF=3.7, Q1 Scopus/WoS, 100 pts**	6, 1

Table 1: Paweł Renc's publications coauthored during doctoral studies at AGH University of Krakow, spanning from 1 October 2021 to 30 September 2025. The three highlighted works represent the core contributions of this PhD thesis, centered on the development of foundation models for electronic health records and federated patient timeline synthesis. The table reports the year, title, venue, and the total number of authors with my authorship position indicated. Blank entries in the authorship column denote collaborative group efforts where the author order does not reflect individual contributions. Only due to limited space not all authors were mentioned - et al. were indicated.

1. Introduction

1.1. Background and Motivation

In recent years, artificial intelligence has emerged as one of the defining technologies of our time, reshaping societies and influencing nearly every aspect of human life. What only a decade ago belonged to the realm of science fiction, machines capable of reasoning, generating, and adapting, has now become a practical reality. Advances in deep learning, particularly the development of transformer architectures and foundation models, have enabled systems capable of performing tasks as varied as real-time translation, creative writing, molecular design, and autonomous navigation.

Healthcare, however, remains one of the greatest challenges for artificial intelligence. Despite steady growth in computational power and algorithmic sophistication, the complexity, sensitivity, and heterogeneity of medical data have limited progress compared to other domains. Costs of care continue to rise worldwide, and in many regions, access to healthcare remains limited. Even in the United States, where nearly eighteen percent of the gross domestic product is devoted to healthcare expenditure [14], outcomes remain suboptimal, with lower life expectancy and higher preventable mortality than in other high-income countries. These systemic inefficiencies underscore the urgent need for new approaches to deliver more efficient, equitable, and proactive care.

The digital transformation of healthcare offers an unprecedented opportunity to address this gap. The widespread adoption of Electronic Health Records (EHRs) has produced a comprehensive digital footprint of patient health over time. These records capture diagnoses, laboratory values, medications, procedures, and increasingly, imaging, clinical notes, and genomic information. Yet, despite their richness, most analytic methods to date treat EHRs in a reductionist manner, collapsing longitudinal trajectories into static snapshots or relying on handcrafted risk scores with fixed thresholds. Such approaches fail to exploit the sequential, heterogeneous, and dynamic nature of patient health.

To unlock the potential of medical data, there is a pressing need for predictive and generative models capable of reasoning across large-scale, episodic, and multimodal health records. Models that can not only forecast outcomes but also simulate plausible future trajectories offer the possibility of a paradigm shift: moving from reactive treatment to proactive and personalized care. This thesis is motivated by the vision of bridging the vast potential of EHR data with the real-world impact of artificial intelligence systems that are trustworthy, scalable, and clinically useful.

1.2. Challenges in Electronic Health Record Modeling

The development of artificial intelligence for healthcare is constrained by several structural challenges inherent to Electronic Health Records (EHRs). First is the question of data availability. Publicly available datasets such as MIMIC [18, 19], eICU [37], and AUMCdb [48] have been instrumental for methodological research; however, they cover only a limited fraction of the data modalities captured in routine care. Access to fully fledged, institution-scale EHRs is typically restricted to those affiliated with healthcare providers and comes with strict regulatory and ethical

oversight. A significant portion of the work in this thesis was done at Massachusetts General Hospital, part of Mass General Brigham, which is a uniquely rich environment for developing and validating models on comprehensive patient data spanning multiple institutions.

Second, traditional early warning systems such as the National Early Warning Score (NEWS) [63] and the Modified Early Warning Score (MEWS) [44] remain widely used; however, they are limited. These tools are designed to provide clinicians with rapid bedside assessments of patient deterioration risk based on a small set of vital signs and simple thresholds. NEWS, for instance, incorporates respiratory rate, oxygen saturation, systolic blood pressure, heart rate, temperature, and level of consciousness into a composite score. Similarly, MEWS uses comparable inputs with fixed cut-offs to stratify patients into low-, medium-, or high-risk categories. While these scores are easy to compute and have become embedded in clinical workflows, their simplicity is also their weakness: they treat patient status as static, ignore the temporal evolution of health trajectories, and cannot adapt dynamically to individual patient contexts. As a result, their predictive accuracy and personalization are limited, especially in complex or rapidly changing clinical situations [42, 21].

Third, classical machine learning methods that followed these early rule-based systems offer some improvements but remain constrained. These models typically depend on preselected predictors engineered from limited data windows, such as the first 24 hours of an admission. This reductionist approach overlooks the richness of patient trajectories, which unfold over time and include complex interdependencies among clinical events. As a result, models trained on narrow slices of data cannot capture the full progression of patient health.

Finally, the problem of generalizability looms large. Healthcare data are highly heterogeneous across institutions due to differences in coding practices, data capture protocols, patient demographics, and workflows. Privacy regulations further limit the sharing of patient data across sites. These factors make it difficult to scale AI solutions beyond the institution where they were originally developed. Without robust frameworks for interoperability and privacy-preserving learning, the promise of AI-enabled healthcare remains restricted to siloed environments, hindering widespread clinical adoption.

Taken together, these challenges highlight the need for a new generation of models that can move beyond static thresholds and narrow prediction windows toward representations capable of capturing full longitudinal trajectories, accommodating heterogeneity, and generalizing across diverse populations. It is precisely this gap that motivated the development of the algorithm presented in this thesis.

1.3. From Early Warning Systems to Foundation Models

The field of clinical risk prediction has undergone a remarkable transformation over the past two decades. Early approaches were dominated by the aforementioned scores, such as NEWS and MEWS. Their success derived from simplicity: clinicians could quickly calculate a score at the bedside and act upon it without computational support. Yet their limitations were equally clear. By reducing patient health to a handful of static variables and fixed thresholds, they ignored the temporal complexity of illness trajectories and the interplay between diverse clinical signals. The rise of data-driven methods marked an important shift. Classical machine learning models, such as logistic regression, random forests, and gradient boosting, incorporate larger sets of predictors and capture nonlinear associations, offering measurable improvements in predictive performance. However, these methods required extensive feature engineering, were confined to narrow temporal windows, and often failed to generalize across sites. They provided incremental advances but stopped short of leveraging the full richness of modern EHR systems.

Deep learning, and transformers in particular, introduce a fundamentally new paradigm. Originally designed for natural language processing, transformers treat data as sequences of tokens, each interpreted in the context of their neighbors. Large language models built on this architecture can learn broad representations of language that

transfer across tasks without retraining. The analogy to healthcare data is powerful: patient histories can likewise be represented as sequences of discrete events, where diagnoses, laboratory results, medications, and procedures act as tokens in a patient health narrative.

The parallels to language are striking. In text, syntax emerges from the order of words, while semantics arises from their meanings and the context in which they appear. Similarly, in healthcare data, temporal ordering provides the syntax: events unfold in a specific sequence, with intervals between them shaping the rhythm of disease progression. The semantics of medicine come from the meaning of each token: an ICD code conveys diagnostic content, a medication token encodes therapeutic action, and a laboratory token reflects physiological state. Just as transformers in language models attend to both syntax and semantics to generate coherent sentences, healthcare transformers can attend jointly to temporal ordering and clinical meaning to generate plausible patient trajectories.

This observation laid the foundation for models that use sequences of health event tokens to model patient health trajectories. By tokenizing the EHR into Patient Health Timelines (PHTs), they allow transformers to learn patterns across the entirety of a patient's history. Once trained, they can generate future Patient Health Timelines in a zero-shot manner, enabling both prediction and simulation without task-specific fine-tuning. This progression, from early warning scores through classical models to transformer-based foundation models, marks a fundamental turning point in the pursuit of scalable, general-purpose artificial intelligence for healthcare.

Hypothesis of the Thesis. The central hypothesis of this dissertation is that tokenized patient health timelines, modeled through generative transformer architectures, provide a universal representation of electronic health records that enables foundation models to achieve scalability, privacy preservation, and clinical fidelity. Specifically, I hypothesize that: (1) zero-shot generative modeling of patient timelines can achieve predictive performance comparable to, or exceeding, specialized task-specific models; (2) adaptive inference over simulated future trajectories can deliver personalized, real-time risk assessments with clinically meaningful explanations; and (3) federated synthesis of synthetic timelines enables cross-institutional training of foundation models without requiring direct sharing of sensitive patient data. Together, these claims define a vision in which a single modeling framework unifies prediction, simulation, explainability, and federated deployment, advancing artificial intelligence toward trustworthy clinical decision support at scale.

1.4. Thesis Contributions and Structure

This thesis is structured around three major contributions that collectively operationalize the above hypothesis.

First, I introduce ETHOS, the Enhanced Transformer for Health Outcome Simulation, which establishes patient health timelines as a tokenized language of care and demonstrates zero-shot trajectory prediction across multiple clinical outcomes.

Second, I extend ETHOS into the Adaptive Risk Estimation System (ARES), which leverages generative forecasting to deliver dynamic, personalized, and explainable risk predictions. ARES illustrates how foundation models can be adapted into clinically actionable decision-support tools.

Third, I present Federated Timeline Synthesis (FTS), which introduces privacy-preserving synthetic patient timelines to enable large-scale, multi-institutional training of foundation models. FTS addresses the challenges of heterogeneity, privacy, and interoperability, demonstrating how foundation models can be trained collaboratively across diverse institutions without sharing raw data.

The final chapters synthesize contributions, reflect on broader implications, and chart future directions toward multimodal integration and patient digital twins.

1.5. Machine Learning in EHR Analytics

Machine learning applied to EHRs has evolved through classical techniques; logistic regression, random forests, gradient boosting, and shallow neural networks. These approaches enabled important clinical prediction tasks but remain limited in their capacity to capture longitudinal structure, generalize across institutions, and adapt to heterogeneous data sources.

Extensive Feature Engineering. Classical models depend heavily on manually engineered features extracted from raw records, a process that is labor-intensive, error-prone, and reliant on domain heuristics.

Static Windows and Limited Temporal Scope. Most models operate on fixed time windows, such as the first twenty-four hours of admission, neglecting the evolving nature of patient health.

Overfitting and Generalizability. Performance often degrades outside the training environment, reflecting overfitting to local practices and coding standards.

Institutional and Population Heterogeneity. Differences in demographics, workflows, and coding limit the portability of models across institutions.

Data Quality Issues. Missingness, irregular sampling, and noisy entries complicate robust modeling.

Ethical and Bias Concerns. Imbalances in training data risk amplifying disparities and embedding bias in predictions.

Interoperability and Privacy Barriers. Privacy regulations and incompatible data systems hinder data sharing, limiting the scale of training.

In summary, while classical machine learning has delivered valuable insights in EHR analytics, its limitations necessitate new paradigms. Foundation models trained on patient health timelines offer a way forward: they reduce reliance on manual feature engineering, embrace longitudinal structure, adapt across heterogeneous populations, and support privacy-preserving collaboration. These innovations collectively motivate the central argument of this thesis: that transformer-based foundation models grounded in patient timelines represent a powerful and generalizable framework for advancing artificial intelligence in healthcare.

2. Related Work

The emergence of foundation models in natural language processing catalyzed a parallel movement in healthcare toward models that learn general representations from large-scale data and transfer across tasks without extensive reengineering. Within electronic health records (EHRs), two broad paradigms have crystallized. The first adapts *natural language* models to clinical corpora, leveraging the maturity of transformer architectures trained on text. The second develops *EHR-native* representations that treat longitudinal patient records as sequences of structured events, often with custom tokenization and temporal encodings. This chapter surveys these trajectories, with attention to methodological choices, scaling behavior, and the persistent tension between expressivity, interoperability, and clinical realism. Where appropriate, it highlights gaps that motivate timeline-centered, generative approaches.

2.1. Natural Language Based EHR Foundation Models

2.1.1. Early Biomedical Language Models

The earliest wave followed the blueprint of domain-adapted language modeling. **BioBERT** extended BERT to large biomedical corpora from PubMed and PubMed Central, demonstrating that pretraining on in-domain text boosts performance on entity recognition and relation extraction [23]. Although not trained on EHRs, BioBERT established the value of massive, domain-specific corpora for downstream clinical tasks.

ClinicalBERT advanced this idea by continuing BERT pretraining on intensive care unit notes from MIMIC-III, thereby capturing the stylistic and semantic idiosyncrasies of clinical documentation [16, 18, 19]. Improvements were observed in readmission prediction and concept extraction, reinforcing a central lesson of representation learning in healthcare, namely that the proximity of pretraining data to deployment data matters.

2.1.2. Scaling Clinical Text Models

Subsequent work emphasized scale, both in tokens and parameters. **GatorTron** exemplified this direction, training an 8.9 billion parameter transformer on over ninety billion words of clinical text drawn from millions of patients [58]. With sufficient diversity and volume, GatorTron reached state of the art performance across concept extraction, relation classification, and natural language inference, suggesting that clinical narratives alone can support broad generalization.

Contemporaneous analyzes amplified this perspective by critiquing the assumptions and limits of text-only clinical models while also documenting their impressive breadth [56]. Collectively, these efforts clarified both the promise of clinical large language models and the liabilities of relying exclusively on narrative text in an ecosystem where much of the salient signal, such as laboratory panels, medication titrations, and procedural details, is recorded in structured or semi-structured form.

2.1.3. Limitations of Natural Language Approaches

Despite strong performance on document-centric tasks, note-based models face three recurring constraints. First, clinical notes are noisy, heterogeneous, and institution specific, which complicates cross-site transfer. Second, narratives incompletely reflect the operational reality captured in orders, codes, and quantitative measurements. Third, free-text timelines are difficult to align with precise event times, complicating causal or counterfactual reasoning. These limitations motivated a shift toward representations that treat the EHR as a *native* sequence of events rather than as a derived text corpus.

2.2. EHR-Native Foundation Models

2.2.1. From Code Embeddings to Contextual Pretraining

Early EHR-native representation learning focused on distributed embeddings for codes and visits. **Med2Vec** learned low-dimensional vectors for medical codes and encounters, enabling similarity queries and simple predictive tasks [8]. **RETAIN** introduced reverse time attention to recover interpretability while modeling sequential visits, highlighting influential events for clinician-facing explanations [9].

Transformers next entered the EHR space. **BEHRT** adapted bidirectional attention to longitudinal records by embedding diagnosis codes with positional and segment encodings that reflect temporal order [26]. **Med-BERT** scaled this paradigm to a vastly larger population, pretraining contextualized embeddings on tens of millions of patient records and transferring them to diverse prediction tasks [38]. **CEHR-BERT** further emphasizes the role of context, showing that BERT-style pretraining over multi-institutional EHRs improves generalization across settings [35]. Together, these studies established that structured EHR tokens, when contextualized at scale, yield robust and transferable representations.

2.2.2. Generative Timeline Modeling

A pivotal turn involved modeling the EHR as an *autoregressive* sequence, moving beyond static classification to sequence generation. **MOTOR** adapted transformers to time-to-event settings by incorporating censored outcomes, bridging survival analysis and sequence modeling to forecast clinical event times [43]. **TransformEHR** framed outcome prediction with an encoder-decoder architecture, emphasizing sequence to sequence dynamics over hand engineered feature sets [59].

Hierarchical designs then appeared. **Hi-BEHRT** extended BEHRT with a hierarchy that captures visit level detail and patient level context simultaneously, improving long range reasoning and performance on complex tasks [25]. In parallel, event stream formalisms, such as **Event Stream GPT (ESGPT)**, advocate modeling irregularly sampled, typed events directly, without aggregating them into fixed visits, thereby preserving fine grained temporal structure [30].

Two empirical cornerstones framed the landscape. **EHRSHOT** offered one of the first systematic evaluations of EHR foundation models in zero shot settings, probing generalization, fairness, and representation choices across tasks and institutions [55]. **Foresight** pretrains an autoregressive transformer on entire patient timelines, demonstrating that generative modeling can yield clinically realistic sequences while supporting multitask transfer [22]. Together, these works consolidated the view that patient trajectories can be modeled analogously to language, with events as tokens and temporal context as syntax.

2.2.3. Synthetic EHR Generation and Autoregressive Data Models

In parallel to predictive modeling, a literature on *synthetic* EHRs matured, driven by privacy, data access, and methodological scalability. **medGAN** pioneered the realistic synthesis of discrete patient records [10], with successors adding temporal fidelity, multimodality, and privacy controls. Privacy oriented frameworks such as **EHR-Safe** seek to balance downstream utility with robust protection from reidentification [60]. More recently, hierarchical autoregressive transformers, including **HALO** and **HiSGT**, modeled timelines as structured token hierarchies, unifying synthesis and prediction in a single language model–style framework [46, 64]. These strands are directly relevant to multi institutional collaboration, where synthetic timelines can serve as intermediaries for training and benchmarking.

2.2.4. Limitations and Open Challenges

Despite notable progress, several challenges remain open. Many approaches still tokenize by visit, which can blur clinically meaningful microdynamics within and across encounters. Integration of heterogeneous modalities, such as imaging, genomics, and free text, is often partial, leaving substantial clinical signals unused. Methods commonly require task specific fine tuning, which trades generality for marginal gains and complicates maintenance at scale. Finally, questions of fairness, calibration, and transportability persist, particularly when models are transferred across institutions with different coding practices, population structures, and care pathways. These limitations have sharpened interest in models that can, in a single framework, represent longitudinal structure, support zero shot inference, and interoperate across sites while preserving privacy.

2.3. Summary

The arc of related work traces a clear trajectory. Language based models have demonstrated the power of scale and domain adaptation on clinical text [23, 16, 58, 56]. EHR native models then embraced the record as a sequence, evolving from code embeddings and attention over visits [8, 9] to contextual transformers trained at population scale [26, 38, 35]. Generative timeline models subsequently reframed prediction as sequence synthesis, enabling richer temporal reasoning and zero shot evaluation [43, 59, 25, 30, 55, 22]. In parallel, synthetic data methods explored privacy preserving routes to collaboration [10, 60, 46, 64]. Together, these strands map a research landscape moving steadily from task specific pipelines toward general purpose EHR foundation models that treat patient trajectories as a language, capture temporal structure explicitly, and aim for interoperable, ethically grounded deployment at scale.

3. Zero-Shot Health Trajectory Prediction using Transformer

This chapter is based on the work published in *npj Digital Medicine* under the title "Zero-Shot Health Trajectory Prediction Using Transformer" [40], where the Enhanced Transformer for Health Outcome Simulation (ETHOS) was first introduced as a foundation model for electronic health records (EHRs). ETHOS demonstrated that Patient Health Timelines (PHTs), when tokenized and modeled using transformer-based generative architectures, can be leveraged to simulate future health trajectories in a zero-shot setting. This capability enables Monte Carlo predictions for a wide variety of clinically relevant tasks, such as mortality risk estimation, readmission prediction, and length-of-stay forecasting—without requiring task-specific labels or retraining.

3.1. Introduction

Healthcare in the United States remains the most expensive worldwide, yet quality and safety outcomes compare unfavorably to those of other developed nations [41]. Although electronic health records (EHRs) are now widely deployed and rule-based decision support systems are commonplace in hospitals, these technologies have not led to consistent improvements in patient outcomes [3]. Artificial intelligence (AI) offers a powerful set of tools to address this gap; however, its integration into clinical practice has been limited. Two fundamental barriers underlie this situation: first, the scarcity of large, well-annotated datasets, which are both costly and labor-intensive to construct; and second, the restricted ability of deployed systems to provide recommendations that are both timely and contextually relevant to the clinician's workflow.

The Enhanced Transformer for Health Outcome Simulation (ETHOS) was designed to directly confront these challenges. ETHOS adapts the transformer deep-learning architecture, originally conceived for natural language processing [50], to the structure of clinical data. Rather than analyzing text, ETHOS processes Patient Health Timelines (PHTs)—comprehensive, chronologically ordered, and tokenized representations of patient health events. Within a PHT, each token corresponds to a discrete unit of information, such as a hospital admission, a prescribed medication, a laboratory test result, or the time interval between consecutive events. By modeling these sequences in an autoregressive manner, ETHOS is capable of generating *future PHTs* (fPHTs), forecasting patient trajectories token by token (Figure 3.1).

ETHOS acquires its generative capabilities entirely through unsupervised pretraining. Once trained, the model can forecast future health events in a zero-shot paradigm without requiring task-specific labeled data or fine-tuning. This establishes ETHOS as a versatile and reusable foundation model for healthcare. Moreover, the design is inherently extensible: with appropriate modifications, ETHOS can incorporate a broad range of clinical data modalities, including but not limited to

- structured EHR codes, such as diagnoses, procedures, and prescriptions,
- laboratory and physiological measurements,
- clinical notes and discharge summaries,

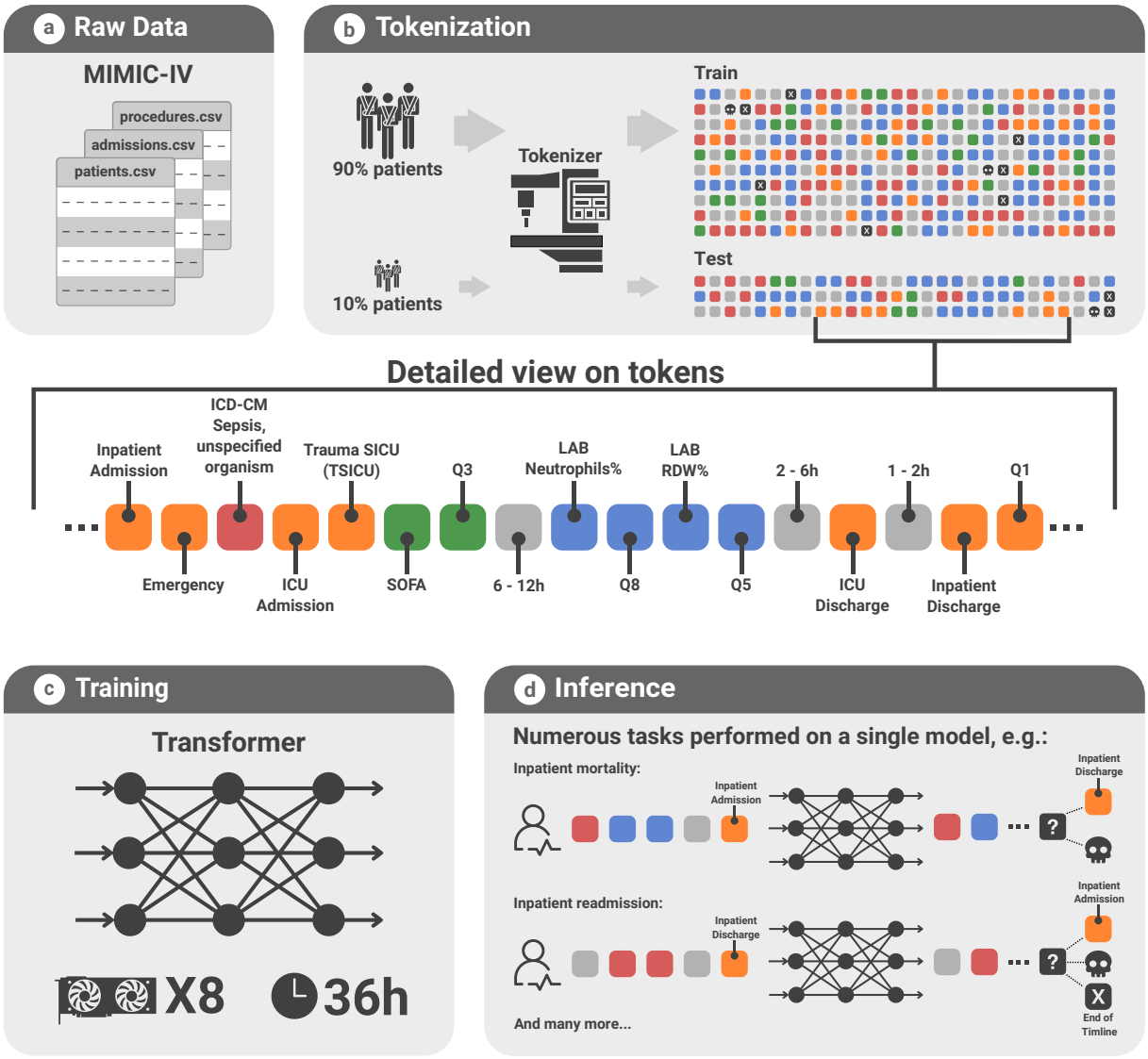


Figure 3.1: The ETHOS workflow: extraction of raw data, tokenization into Patient Health Timelines (PHTs), autoregressive training of a transformer model, and zero-shot inference for diverse downstream clinical tasks.

– imaging data from radiology and pathology,

– continuous monitoring signals from ICUs or wearable devices, and

– high-dimensional omics data, such as genomics and proteomics.

To establish feasibility, ETHOS was trained and evaluated using the MIMIC-IV v2.2 dataset [18, 19], a large-scale, publicly available collection of critical care records. Importantly, this dataset was utilized in its original form, without data cleaning, imputation, or removal of inconsistent entries. Common inconsistencies, such as discharge dates preceding admission dates, were left unaltered. The guiding assumption was that, with sufficiently large datasets and an appropriately designed tokenization and training framework, ETHOS would learn to tolerate noise and irregularities. This approach acknowledges the reality that healthcare data are inherently imperfect: errors, missing values, and inconsistencies are not only common but also often unavoidable. Manual data curation at scale is impractical and risks introducing additional biases. Demonstrating resilience to such noise is, therefore, critical for the deployment of robust and generalizable healthcare AI.

A defining characteristic of ETHOS, which distinguishes it from many existing approaches (see Chapter 2), is that no task-specific supervision is required. Once pretrained, ETHOS can simulate diverse possible health trajectories and derive predictions from these generative simulations. This eliminates the need for repeated retraining or curated labels for each new clinical question. Furthermore, ETHOS does not impose restrictive inclusion criteria for patient data, supporting scalability to datasets that encompass millions of patients. In this way, ETHOS represents a step toward a unified and general-purpose foundation model for healthcare.

3.2. Methods

3.2.1. Data

The empirical basis of this research is the Medical Information Mart for Intensive Care (MIMIC-IV) database, version 2.2 [18, 19]. This resource, created through collaboration between the Massachusetts Institute of Technology, Beth Israel Deaconess Medical Center, and Philips Healthcare, is one of the most comprehensive open-access repositories of de-identified clinical data. MIMIC-IV includes detailed information on more than 200,000 patients admitted to hospitals and intensive care units at BIDMC in Boston, Massachusetts, between 2008 and 2019. Its breadth and diversity make it particularly well suited for the development and evaluation of large-scale foundation models in healthcare.

To construct Patient Health Timelines (PHTs), data were extracted from a broad selection of MIMIC-IV tables to ensure comprehensive coverage of patient trajectories. The following tables were incorporated:

1. `patients`: static attributes such as gender, date of birth, and date of death,
2. `admissions`: admission and discharge information and contextual metadata,
3. `icustays`: details of ICU admissions, including start and end times and type of unit,
4. `labevents`: laboratory test results, restricted to the 200 most frequent tests that cover approximately 95% of completed measurements,
5. `procedures`: procedures coded in ICD-10-PCS,
6. `diagnoses`: diagnostic codes primarily in ICD-10-CM, with ICD-9 converted into ICD-10-CM using official conversion tables[1],
7. `emar`: electronic medication administration records, mapped to the Anatomical Therapeutic Chemical (ATC) classification system[2], with Generic Sequence Numbers converted using published mappings [5],
8. `omr`: outpatient measurements such as blood pressure or BMI,
9. `services`: the clinical service responsible for the patient during hospitalization,
10. `drgcodes`: Diagnosis-Related Group assignments for hospitalization-based resource classification, and
11. `sofa`: Sequential Organ Failure Assessment scores from derived MIMIC tables.

Tables not listed above, particularly those containing unstructured free-text clinical notes, were excluded. While potentially informative, these sources require additional natural language processing pipelines to transform text into structured tokens. Excluding them at this stage allowed for the development of a tokenization framework optimized for structured and coded data.

[1] `https://www.cms.gov/medicare/coding-billing/icd-10-codes`
[2] `https://www.whocc.no`

3.2.2. Patient Health Timelines and Tokenization

The cornerstone of ETHOS is the design of *Patient Health Timelines* (PHTs), which recast the complete trajectory of a patient's medical history into a symbolic sequence suitable for autoregressive modeling. This representation provides a unifying grammar for heterogeneous EHR data, enabling transformer architectures to parse and generate longitudinal clinical information. PHTs integrate dynamic medical events—diagnoses, medications, procedures, laboratory measurements, and vital signs—along with static demographic descriptors, producing a format that is both comprehensive and temporally coherent (Figure 3.1).

Chronological assembly of events. The construction of a PHT begins with the alignment of events drawn from multiple MIMIC-IV tables into a strictly ordered sequence. Each event is timestamped as the patient age, stored in 64-bit floating-point precision with six significant digits. This removes the dependency on absolute dates, which are often masked for privacy, while retaining precise temporal ordering relative to the patient's lifetime. Once ordered, each event is transformed into one or more tokens under the ETHOS scheme. Simple events may require a single token, while complex entries, such as multi-component lab results or procedures, can expand into as many as seven tokens, ensuring the preservation of clinically relevant details. All tokens derived from a single event share the same timestamp, maintaining temporal integrity at the event level.

Encoding temporal gaps. Because EHR data is irregularly sampled, ETHOS explicitly models the elapsed time between events using interval tokens (Figure 3.2). Thirteen such tokens capture durations spanning from minutes to months. When the gap between events is shorter than five minutes (the smallest supported unit), no token is added. For longer gaps, a compositional strategy is employed: durations exceeding one year are represented by concatenating multiple six-month tokens. For example, a 1.4-year interval is expressed as three six-month tokens, while a 1.76-year interval requires four. This method embeds temporality directly into the symbolic representation, replacing absolute timestamps with interpretable relative markers while preserving continuity in the token stream.

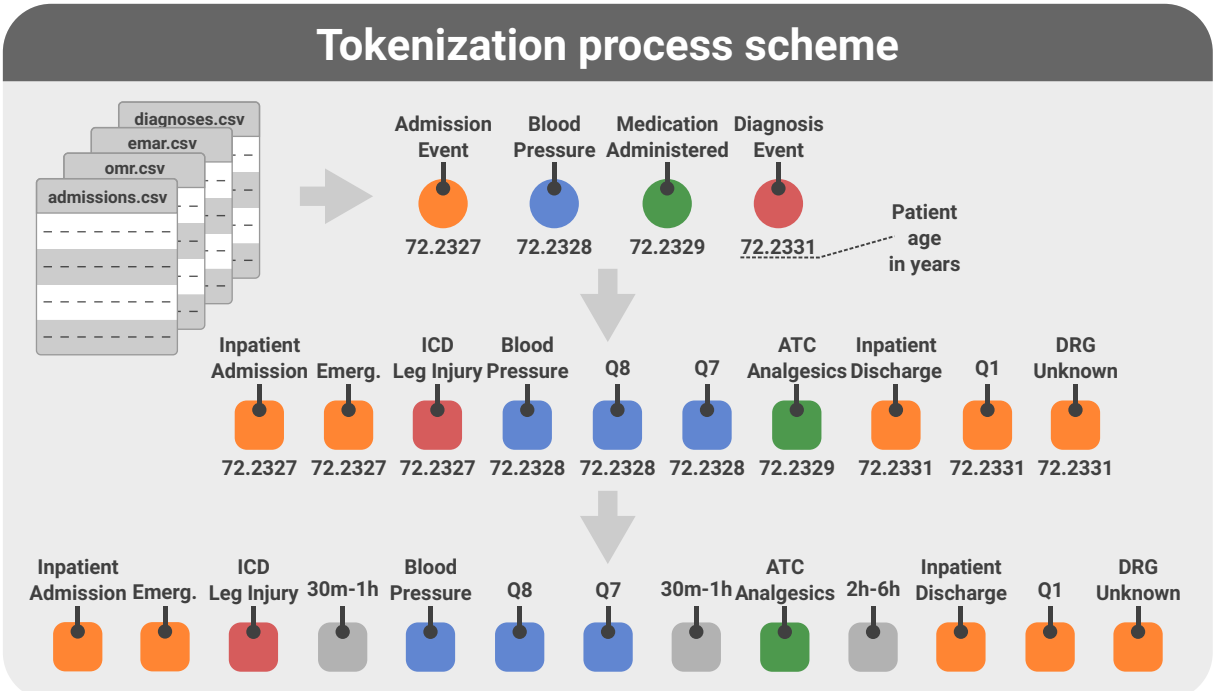


Figure 3.2: Stages of PHT construction and tokenization. Raw chronological EHR events are transformed into tokens, enriched with explicit time-interval markers, and concatenated into unified patient timelines.

Static patient attributes. PHTs also incorporate invariant patient descriptors as tokens to provide baseline context. Gender, marital status, race, body mass index (BMI), birth date, and the start date of the timeline are represented as fixed tokens. Although such variables may change over time in reality, they are treated as constants in the MIMIC-IV dataset and are encoded accordingly. To guaranty their inclusion in every inference, the first six positions of each 2,048-token context window are reserved for these attributes. The sixth token also encodes the temporal position of the seventh, which marks the first dynamic event. This strategy consistently embeds demographic context while minimizing architectural complexity.

Encoding continuous variables. Continuous values, such as laboratory concentrations or vital signs, are discretized into quantile-based tokens. Histograms were computed over the training data for each measurement type, and values were mapped into one of ten bins (Q1–Q10). This design balances precision with interpretability: clinically significant changes are typically reflected in decile shifts rather than subtle fluctuations. Embedding analyzes confirmed that ETHOS internalized ordinal structure, with adjacent quantiles occupying contiguous regions in latent space (Figure 3.5b).

Domain-specific tokens. Certain variables are tied to particular points in the care process and are therefore inserted in specific contexts. DRG codes, which are defined at discharge, are positioned after a discharge token, a quantile representing length of stay, and a discharge destination token. Similarly, Sequential Organ Failure Assessment (SOFA) scores are added immediately after ICU admission tokens, together with ICU type tokens. During inference, ETHOS does not access these true values; rather, it autoregressively predicts them from prior context, ensuring causal validity while leveraging ground-truth values during training.

Hierarchical coding systems. To expose the semantic structure, ETHOS decomposes complex medical codes into multiple tokens (Figure 3.4). For ATC drug codes, up to three tokens are used: the first encodes the initial three characters, the second encodes the subsequent character, and an optional third encodes any suffix. ICD-10-CM diagnoses are split into three tokens, while ICD-10-PCS procedures, consisting of seven alphanumeric characters, are represented character-by-character. This hierarchical tokenization mirrors the semantics of these systems, where earlier characters denote broad categories and later characters add specificity. Learned embeddings demonstrated that ETHOS organized these tokens into clusters consistent with medical ontologies, validating the approach.

Unified symbolic representation. The final outcome of this design is a flattened chronological sequence that integrates demographic context, categorical medical codes, quantile-discretized measurements, and explicit temporal markers into a unified symbolic representation. Unlike visit-based or tabularized methods, which fragment patient histories, ETHOS PHTs preserve continuity and fine-grained temporal dynamics. This sequential representation allows autoregressive transformers to exploit dependencies across long clinical trajectories, analogously to natural language. Figure 3.5 illustrates insights from this scheme: (a) the distribution of token categories across the dataset, (b) the embedding geometry of quantile tokens, and (c) the embedding organization of time-interval tokens, which align with their real-world semantics. These results confirm that ETHOS tokenization yields structured, interpretable, and clinically meaningful embeddings.

3.2.3. ETHOS Training

ETHOS employs a decoder-only transformer architecture closely aligned with the design of generative pre-trained transformers (GPT) [50]. Implementation was based on the NanoGPT codebase[3], modified to incorporate

[3] `https://github.com/karpathy/nanoGPT`

learnable positional embeddings rather than fixed sinusoidal encodings. Patient timelines were concatenated into a single long token stream, each terminated by an explicit `End of timeline` marker. The training objective was autoregressive next-token prediction under causal masking, identical to large-scale language modeling.

Given the size of the dataset and the complexity of the model, optimization required resources on the order of those used to train GPT-2 [6]. Hyperparameters were initially set following GPT-2 conventions and then tuned heuristically to minimize validation loss (Figure 3.6). Training procedures adhered to established best practices for transformer optimization [50, 47].

The resulting embeddings captured semantically coherent structures. Quantile tokens were arranged in ordinal order, time-interval tokens aligned proportionally to their durations, and categorical code tokens clustered by hierarchy. Supplementary Figures 3 and 4 illustrate these relationships. These findings suggest that ETHOS internalized meaningful latent structures directly through unsupervised training. To facilitate reproducibility, the full implementation and training scripts have been released publicly at `https://github.com/ipolharvard/ethos-paper`.

3.2.4. Evaluation of Clinical Outcomes and Tasks Using ETHOS

The evaluation of ETHOS was structured around a set of downstream clinical prediction tasks, carefully selected to both demonstrate the breadth of the model's zero-shot capabilities and enable comparisons with prior work on the MIMIC dataset. Patients were randomly partitioned into training and testing cohorts in a 90%/10% ratio, as summarized in Table 3.1. The test cohort was subsequently used to assess ETHOS across multiple outcome types, including mortality prediction, length of stay estimation, readmission risk assessment, SOFA score regression, and Diagnosis-Related Group (DRG) classification.

Inpatient and ICU mortality. For hospital mortality, ETHOS estimated the probability of death at the time of admission. The generative process was initiated by seeding the model with an admission token and allowing it to autoregressively generate until either a discharge or death token was produced. Each simulation was repeated 20 times. The mortality probability was then computed as $N/20$, where N is the number of simulated trajectories that terminated in death before discharge. ICU mortality was evaluated analogously, with two variants: one beginning at ICU admission and another starting 24 hours after ICU admission, reflecting distinct clinical decision points. In the same simulated timelines, the length of stay in the ICU was calculated by summing the time-interval tokens up to discharge. Simulated patients who died in the ICU were excluded from LOS calculations, ensuring that estimates reflected survivors only. Across these experiments, 20 repetitions yielded 21 unique probability estimators, sufficient for the robust construction of ROC curves that fit closely to Gaussian models (Figure 3.8).

Readmission. Thirty-day inpatient readmission was modeled by initiating the generative process at the discharge token of each hospitalization. Sequences were extended until one of three outcomes occurred: (1) a new admission token, (2) a death token, or (3) cumulative simulated time tokens exceeding 30 days. Each case was simulated 20 times, and the proportion of runs terminating with a new admission token within 30 days was reported as the estimated probability of readmission.

SOFA score regression. ETHOS was also evaluated on severity scoring using the Sequential Organ Failure Assessment (SOFA). For each ICU admission, the timeline was constructed to include an ICU admission token, a SOFA token, and a quantile token representing the observed first-day SOFA score. During inference, ETHOS predicted SOFA by generating probability distributions over quantile tokens. These probabilities were mapped to the expected value associated with each quantile, providing a continuous estimate of SOFA (Figure 3.9a). Although

Characteristics	Train/Validation	Test	Total
Patient number	241,015	26,758	267,773
Age, years, mean (std)	50.27 (20.76)	50.15 (20.80)	50.25 (20.77)
Gender			
Female	130,115	14,473	144,588
Male	110,900	12,285	123,185
Race			
White	109,274	12,090	121,364
Unknown	87,420	9,724	97,144
Black	21,048	2,361	23,409
Hispanic	8,991	1,014	10,005
Other	7,506	823	8,329
Asian	6,776	746	7,522
Marital Status			
Unknown	84,533	9,369	93,902
Married	70,297	7,649	77,946
Single	60,822	6,910	67,732
Widowed	15,068	1,719	16,787
Divorced	10,295	1,111	11,406

Table 3.1: Demographic characteristics of the dataset. Characteristics are reported at the time of the first hospital admission and were used by ETHOS in constructing Patient Health Timelines.

this design introduces an apparent causality violation at training time—since first-day SOFA scores are only known retrospectively—the true values were not used during inference. Instead, ETHOS generated SOFA predictions solely from prior information, thereby maintaining causal validity in predictive settings.

DRG classification. Finally, ETHOS was applied to the task of DRG classification, which represents both an administrative and a clinical outcome. DRG tokens were consistently placed after discharge tokens and a quantile token encoding the length of stay. During inference, ETHOS produced probability distributions across 771 DRG classes. Performance was evaluated using top-k accuracy, with both top-1 and top-2 reported (Figure 3.9c). This experiment underscored the model's ability to generalize beyond direct clinical outcomes to tasks critical for hospital operations and reimbursement.

3.2.5. Statistical Analysis

The performance of ETHOS across this broad set of outcomes was quantified using metrics appropriate to the type of endpoint. For binary classification tasks, including inpatient and ICU mortality as well as readmission, Receiver Operating Characteristic (ROC) analysis was performed. Empirical ROC points were fitted with Gaussian models, allowing for unequal variances under each hypothesis, which yielded smooth ROC curves. Areas under the curve (AUCs) were reported as summary statistics, with 95% confidence intervals estimated through bootstrapping.

For multiclass classification tasks such as DRG assignment, top-1 and top-2 accuracies are reported, capturing both exact and near-exact performance. Regression tasks, including SOFA score estimation and ICU length of stay,

were evaluated using the mean absolute error (MAE). As with AUCs, 95% confidence intervals for MAE were obtained using bootstrap resampling. All analyzes were conducted in Python using `numpy` and `scikit-learn`. To ensure transparency, the entire analysis pipeline was made available alongside the ETHOS codebase.

3.2.6. Comparison of ETHOS to Existing Methods

To situate ETHOS within the broader methodological landscape, I compared it with both traditional machine learning approaches and modern large language models. These baselines were chosen to represent, respectively, established paradigms for EHR prediction and the emerging class of general-purpose foundation models. The methodological details of these comparisons are described below.

Traditional and graph-based methods. The principal benchmark task was the 30-day hospital readmission prediction, a long-standing testbed in health informatics. As a baseline, I followed the feature extraction methodology of Tang et al. [45], which derives patient-level features from hospitalization records. Their full approach relies on a spatiotemporal graph neural network (STGNN) that constructs graphs by computing pairwise similarities across admissions. While effective on their dataset of 14,500 admissions, this procedure proved computationally infeasible at the scale of MIMIC-IV, which contains over 400,000 admissions. The pairwise similarity computation required for graph construction exceeded memory capacity, even on a high-performance node with 2 TB of RAM.

Consequently, I restricted replication to the preprocessing and feature engineering pipeline, which was faithfully re-implemented with adapted code from the authors' repository[4]. On these features, I trained conventional baselines, including logistic regression and XGBoost. Because these models lack temporal awareness, temporal variation was compressed into descriptive statistics: for each time-varying variable, the minimum, first quartile, median, third quartile, and maximum were included as features. Additionally, the admission day was encoded explicitly to retain some temporal context. This design allowed static models to approximate longitudinal dynamics within their inherent methodological limits.

Comparison with large language models. ETHOS was also benchmarked against GPT-4o, a state-of-the-art proprietary large language model. While GPT-4o was not trained on EHR data, its scale and general reasoning abilities make it an important comparator. To ensure fairness, I designed a structured prompt pipeline that supplied GPT-4 with the same patient information available to ETHOS. Each prompt contained: (1) task instructions, (2) static patient attributes, (3) a tokenized patient timeline truncated to 2,048 tokens, and (4) explanatory notes describing token subgroups. GPT-4o was then asked to estimate 30-day readmission probabilities for 2,000 test patients. Two decoding temperatures (0.3 and 0.5) were evaluated to examine the impact of stochasticity. The full prompt design and experimental notebook are included in the ETHOS repository[5].

Although GPT-4o represents one of the most powerful general-purpose LLMs available, it was not optimized for temporal EHR sequences. ETHOS, in contrast, was explicitly designed to treat patient records as symbolic temporal sequences with explicit interval tokens, allowing it to autoregressively simulate future trajectories. This distinction proved decisive in comparative performance, as discussed in the results section below.

3.3. Results

ETHOS was comprehensively evaluated across a diverse set of downstream tasks designed to reflect clinically meaningful outcomes derivable from EHRs. Importantly, ETHOS was assessed in a zero-shot regime, meaning

[4] `https://github.com/ipolharvard/readmit-stgnn`
[5] `https://github.com/ipolharvard/ethos-paper/blob/master/notebooks/llm_readmission_task.ipynb`

that no task-specific retraining or fine-tuning was required. Tasks were grouped into four broad categories: (1) patient-level outcomes, including inpatient and ICU mortality; (2) healthcare utilization outcomes, such as length of stay (LoS); (3) severity estimation tasks, exemplified by Sequential Organ Failure Assessment (SOFA) score prediction; and (4) administrative outcomes, such as Diagnosis-Related Group (DRG) classification. In addition, readmission was evaluated both at the hospital and ICU levels. Collectively, these tasks span binary classification, regression, and multiclass classification, providing a comprehensive demonstration of ETHOS's versatility as a foundation model.

3.3.1. Readmission Benchmark Results

Figure 3.7 summarizes the comparative results for the 30-day hospital readmission task. Logistic regression and XGBoost baselines achieved AUCs of 0.650 and 0.675, respectively, while recurrent models such as RNNs and LSTMs reached 0.661 and 0.660. ETHOS substantially outperformed all static and sequential baselines, achieving an AUC of 0.749 (95% CI: 0.742–0.745).

When compared against GPT-4o, ETHOS also exhibited superior predictive performance. GPT-4o achieved AUCs of 0.622 (temperature 0.3) and 0.632 (temperature 0.5), substantially below ETHOS. Despite its general-purpose reasoning capacity, GPT-4o's performance was closer to simple baselines than to ETHOS, emphasizing the benefits of temporal tokenization and autoregressive simulation in clinical prediction. By treating patient histories as structured symbolic sequences enriched with explicit time-interval tokens, ETHOS was able to generate more accurate and well-calibrated risk estimates.

3.3.2. Mortality Prediction

Mortality prediction remains a cornerstone benchmark for evaluating EHR-based models, serving as a stringent test of discriminatory power. ETHOS achieved excellent results in both inpatient and ICU-specific mortality tasks. For inpatient mortality, ETHOS attained an AUC of 0.921 (95% CI: 0.908–0.931), while ICU mortality prediction reached an AUC of 0.927 (95% CI: 0.914–0.938). These values place ETHOS at or above the leading baselines reported in the literature. For example, Pang et al. achieved an ICU mortality AUC of 0.918 using XGBoost [36], whereas less sophisticated deep learning models have been reported at substantially lower levels (AUC $\approx$ 0.64 [7]).

Performance was also robust in clinically challenging subgroups. For patients with sepsis, ETHOS achieved an AUC of 0.889 (95% CI: 0.870–0.906), surpassing the discriminatory ability of SOFA-based severity scores, which typically plateau at approximately 0.76 [34]. These findings highlight ETHOS's ability to capture subtle temporal dependencies and causal pathways in longitudinal data that static indices cannot fully reflect.

3.3.3. Length of Stay Prediction

Length of stay (LoS) is a critical metric for hospital resource allocation and operational planning. ETHOS predicted ICU LoS with a mean absolute error (MAE) of 2.26 days (95% CI: 2.16–2.35). This performance is on par with, or slightly better than, specialized models trained explicitly for LoS prediction, which typically report MAEs around 2.4 days [7].

Beyond point accuracy, ETHOS provides additional value by generating distributions of LoS through repeated autoregressive sampling of future timelines. This scenario-based approach enables uncertainty quantification, offering not just a single prediction but a range of plausible outcomes. Such probabilistic estimates are particularly valuable in hospital operations, where planning must account for variability and resource constraints.

3.3.4. SOFA Score Estimation

ETHOS demonstrated the capacity to approximate established clinical scoring systems in a zero-shot manner. For first-day Sequential Organ Failure Assessment (SOFA) scores, ETHOS achieved a MAE of 1.50 (95% CI: 1.47–1.53). This result represents, to our knowledge, the first demonstration of zero-shot SOFA prediction from admission data alone.

The ability to estimate severity indices directly from longitudinal records without explicit labels has important implications. It suggests that ETHOS can internalize scoring logic from patient trajectories, effectively rediscovering structured clinical indices from raw data. This property could be particularly beneficial in settings where SOFA or similar scores are not consistently recorded, providing a scalable surrogate for clinical decision support.

3.3.5. DRG Classification

Diagnosis-Related Group (DRG) classification was used to evaluate ETHOS on a large-scale multiclass prediction problem involving 771 categories. At discharge, ETHOS achieved a top-1 accuracy of 84.8% (95% CI: 84.4–85.2), a dramatic improvement over the 52% accuracy reported by Wang et al. using discharge-note-based LLMs [51].

This substantial improvement reflects ETHOS's ability to exploit the full span of longitudinal patient trajectories, incorporating diagnostic, procedural, and therapeutic events rather than relying on a single text document. Because DRG codes play a central role in hospital reimbursement systems, improved accuracy in this task has direct implications for both administrative efficiency and healthcare financing.

3.3.6. Summary of Downstream Evaluation

Across all tasks, ETHOS demonstrated strong, often state-of-the-art performance. Mortality prediction matched or exceeded published baselines, with particularly strong results in high-variability cohorts such as sepsis. Length of stay prediction equaled specialized models while uniquely providing uncertainty estimates through generative sampling. Readmission prediction approached the performance of heavily engineered graph-based methods despite operating without task-specific feature engineering. SOFA score estimation showcased ETHOS's ability to reconstruct structured severity indices in a zero-shot manner, and DRG classification advanced the state of the art by a wide margin.

Together, these findings establish ETHOS as a versatile foundation model for healthcare. Its zero-shot generative framework allows a single pretrained model to support a wide spectrum of predictive and administrative tasks, reinforcing patient timeline tokenization as a unifying paradigm for healthcare AI.

3.4. Discussion: Robustness and Limitations

A central observation from the development and evaluation of ETHOS is its robustness in the face of the noisy and imperfect character of real-world EHR data. Unlike many prior studies that rely on extensive preprocessing, imputation of missing values, or exclusion of anomalous records, ETHOS was trained directly on raw MIMIC-IV data. This dataset is known to contain numerous inconsistencies, including missing values, irregular sampling intervals, heterogeneous coding practices across admissions, and even implausible entries, such as discharge dates preceding admission dates. Despite these imperfections, ETHOS consistently delivered strong predictive performance across mortality, severity scoring, length of stay, readmission, and administrative classification tasks. These findings demonstrate that large transformer-based foundation models are capable of learning to accommodate, rather than eliminate, the variability inherent in clinical data. This property is essential for real-world deployment

since clinical datasets at scale will inevitably include noise, and efforts to enforce perfect data quality are both impractical and potentially bias-inducing.

Beyond resilience to noise, ETHOS exhibited internal representations that reflect clinically coherent structures. Figure 3.10 illustrates how token embeddings, when projected into low-dimensional spaces, align with human-understandable categories. For example, ICD-10-CM diagnostic tokens cluster by broad diagnostic areas, even though ETHOS was never explicitly trained to recognize such groups. Similarly, quantile tokens representing patient age and PHT start date are arranged in ordinal order, with small distances between mid-range values (Q4–Q6) and wider spacing at the extremes (Q9–Q10). This mirrors clinical reasoning, where modest fluctuations within the normal range are less consequential than extreme deviations. Time-interval tokens also demonstrate alignment with actual durations, further highlighting that ETHOS internalized temporal structure directly from raw training sequences. Together, these patterns show that ETHOS not only predicts outcomes accurately but also develops semantically and temporally meaningful latent spaces, providing an interpretable substrate for clinical reasoning.

Despite these strengths, ETHOS, in its current form, has several important limitations. A primary constraint is the capped context length of 2,048 tokens. While this is sufficient for many hospitalizations, long or complex trajectories may exceed this limit, forcing the truncation of clinically relevant events. As a consequence, important information late in a timeline may not be available to the model during inference. Overcoming this limitation will require advances in architecture or training methodology, such as long-context transformers, memory-augmented mechanisms, or recurrent compression strategies that preserve salient events while reducing sequence length.

Another limitation lies in the range of modalities incorporated. ETHOS version 1 was restricted to structured EHR data, encompassing diagnoses, laboratory results, medications, procedures, demographics, and vital signs. Other modalities central to modern clinical practice were not yet integrated, including unstructured clinical notes, radiology and pathology images, continuous monitoring data from ICUs, and high-dimensional omics datasets. Integrating these sources will be necessary to move toward a comprehensive digital representation of patient health. ETHOS's tokenization strategy provides a clear pathway for such integration; however, realizing this potential requires significant future work in multimodal fusion.

Scalability is both a strength and a challenge. ETHOS was trained on hundreds of thousands of patient trajectories, and its performance is expected to improve as the training corpus grows to millions or even tens of millions of patients. However, the computational requirements for training and inference are substantial. While comparable to those of language models of similar scale, further optimization of training efficiency and inference latency will be necessary for practical deployment in hospital environments where near real-time predictions are required.

Finally, issues of fairness, generalizability, and institutional bias remain open. Although ETHOS demonstrated resilience to noise in MIMIC-IV, the dataset is derived from a single academic medical center, which may limit its representativeness across different health systems, populations, or countries. Extending ETHOS to diverse datasets and rigorously evaluating performance across demographic subgroups will be essential to ensure equitable outcomes. Addressing these challenges will likely involve federated training paradigms and fairness-aware evaluation, which are discussed in later chapters.

In summary, ETHOS exhibits notable robustness to noisy input data and develops clinically meaningful latent representations without explicit supervision. At the same time, limitations regarding sequence length, modality scope, computational cost, and generalizability highlight important areas for refinement. These constraints do not diminish the central finding: a single foundation model can operate effectively across diverse prediction tasks in real-world EHRs without task-specific retraining. Instead, they point the way toward future research directions necessary to translate ETHOS from proof-of-concept into a truly comprehensive and deployable clinical foundation model.

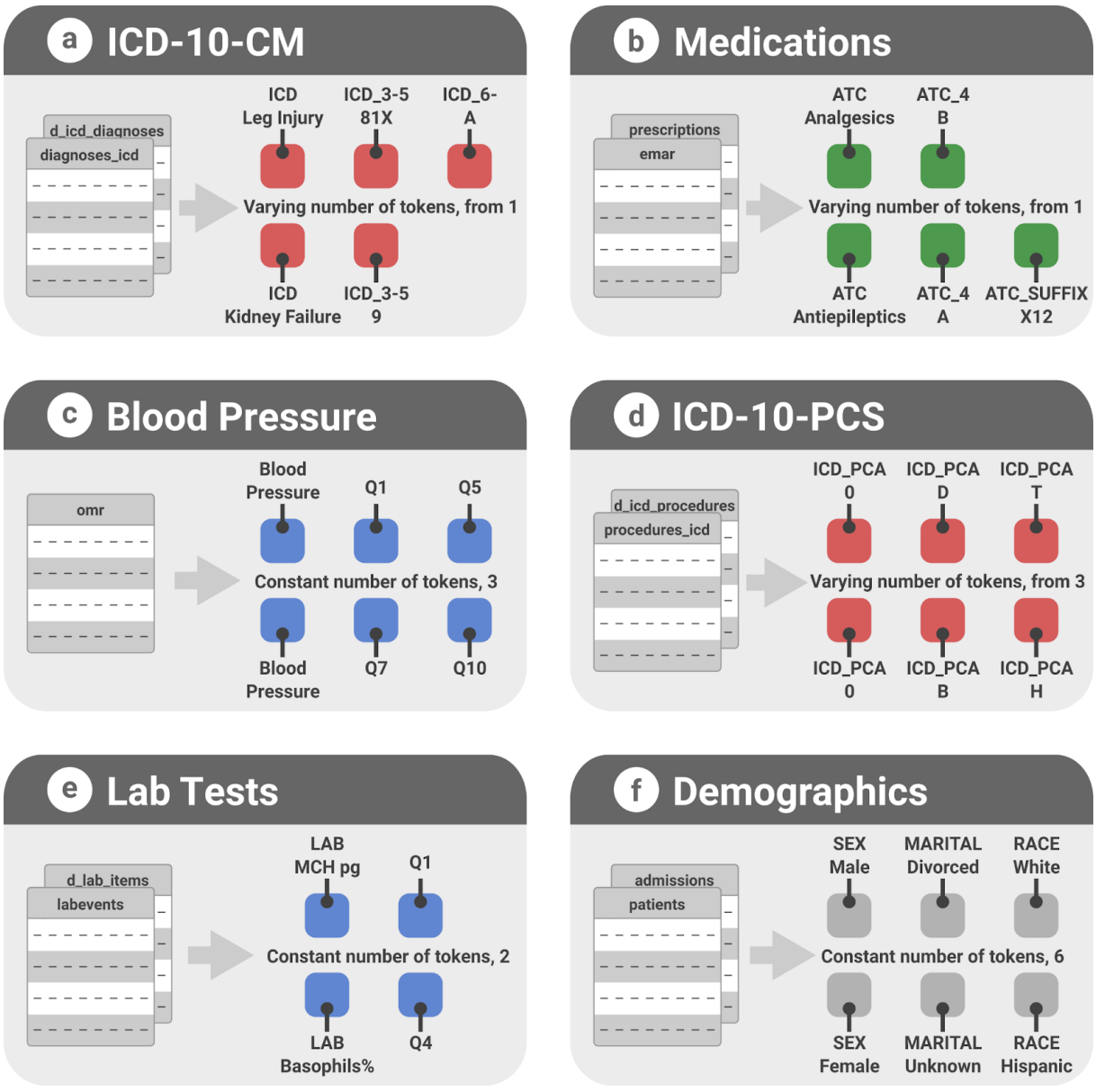


Figure 3.3: Examples of ETHOS tokenization across multiple modalities, including diagnoses, procedures, medications, laboratory tests, vital signs, and demographic attributes.

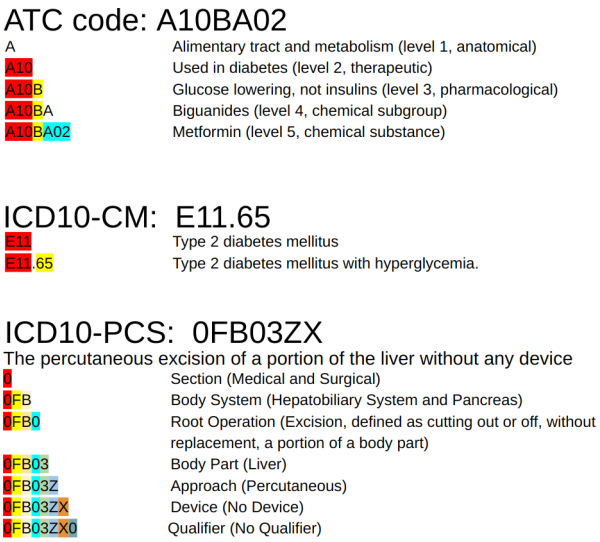


Figure 3.4: Hierarchical decomposition of ATC and ICD coding systems. ETHOS leverages multi-token representations to expose meaningful structure to transformer attention mechanisms.

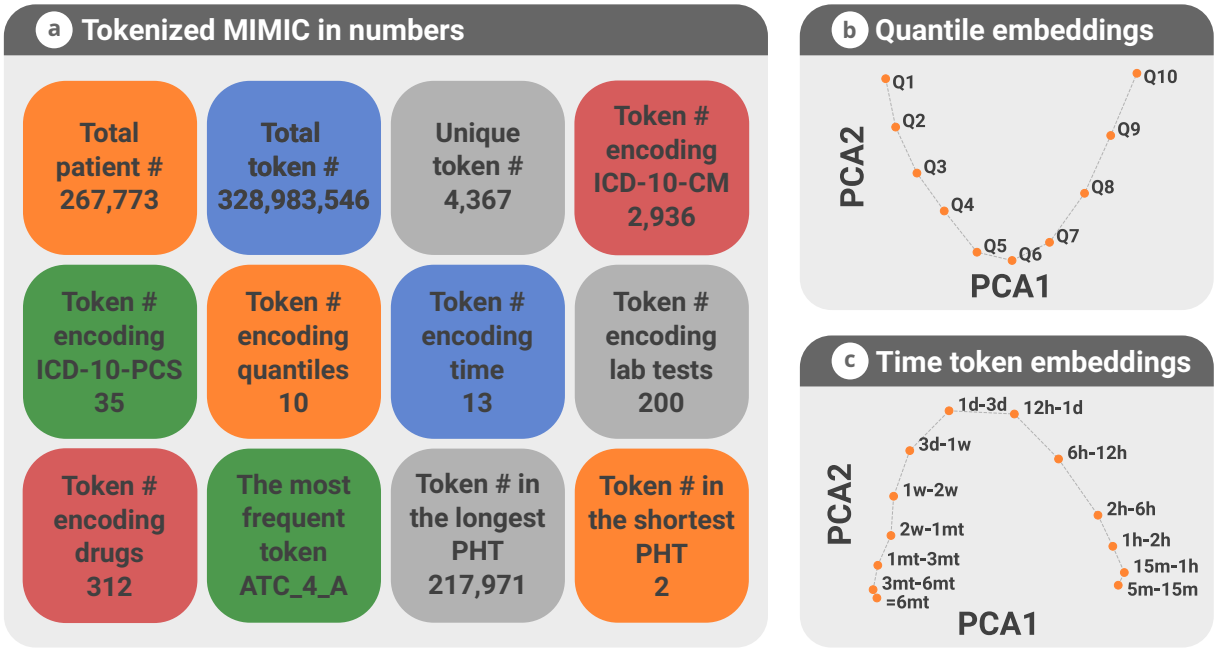
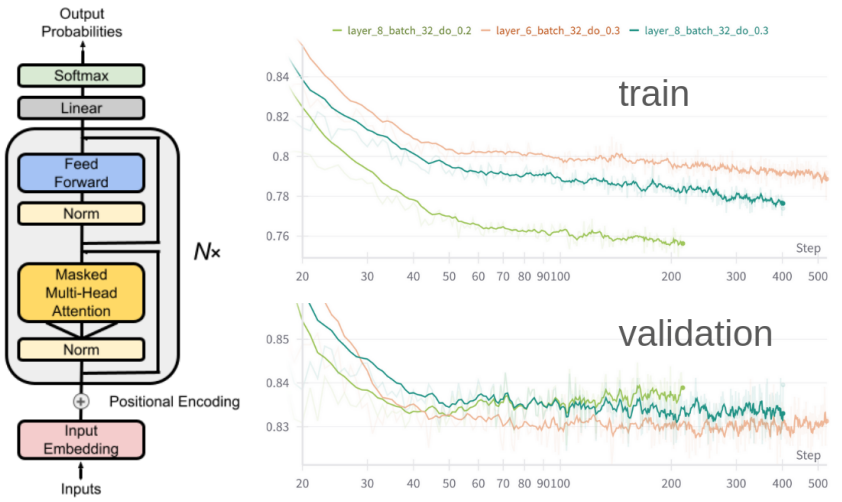


Figure 3.5: Tokenization and embedding visualizations of MIMIC-IV data. (a) Distribution of token types. (b) Embedding geometry of quantile tokens. (c) Embedding geometry of time-interval tokens.



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

Figure 3.6: Architecture of the Transformer Decoder Model employed in ETHOS. Included are select training traces and a detailed account of the operational model, complete with parameter specifications outlined in the accompanying table.

Method	AUC	95% CI
Logistic regression	0.650	0.643-0.656
Xgboost	0.675	0.669-0.681
RNN	0.661	0.654-0.668
LSTM	0.660	0.654-0.667
ETHOS (ours)	0.749	0.742-0.745

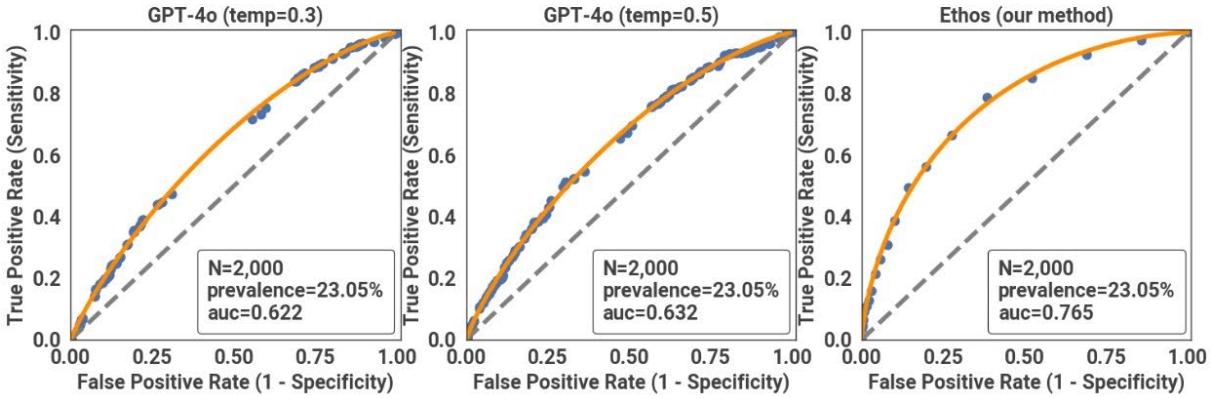


Figure 3.7: **Comparative performance on 30-day readmission.** Top: AUC values with 95% confidence intervals for logistic regression, XGBoost, RNN, LSTM, and ETHOS. Bottom: ROC curves for GPT-4o (temperature 0.3 and 0.5) and ETHOS on 2,000 test patients (prevalence 23.05%). ETHOS consistently outperformed both traditional baselines and the general-purpose LLM.

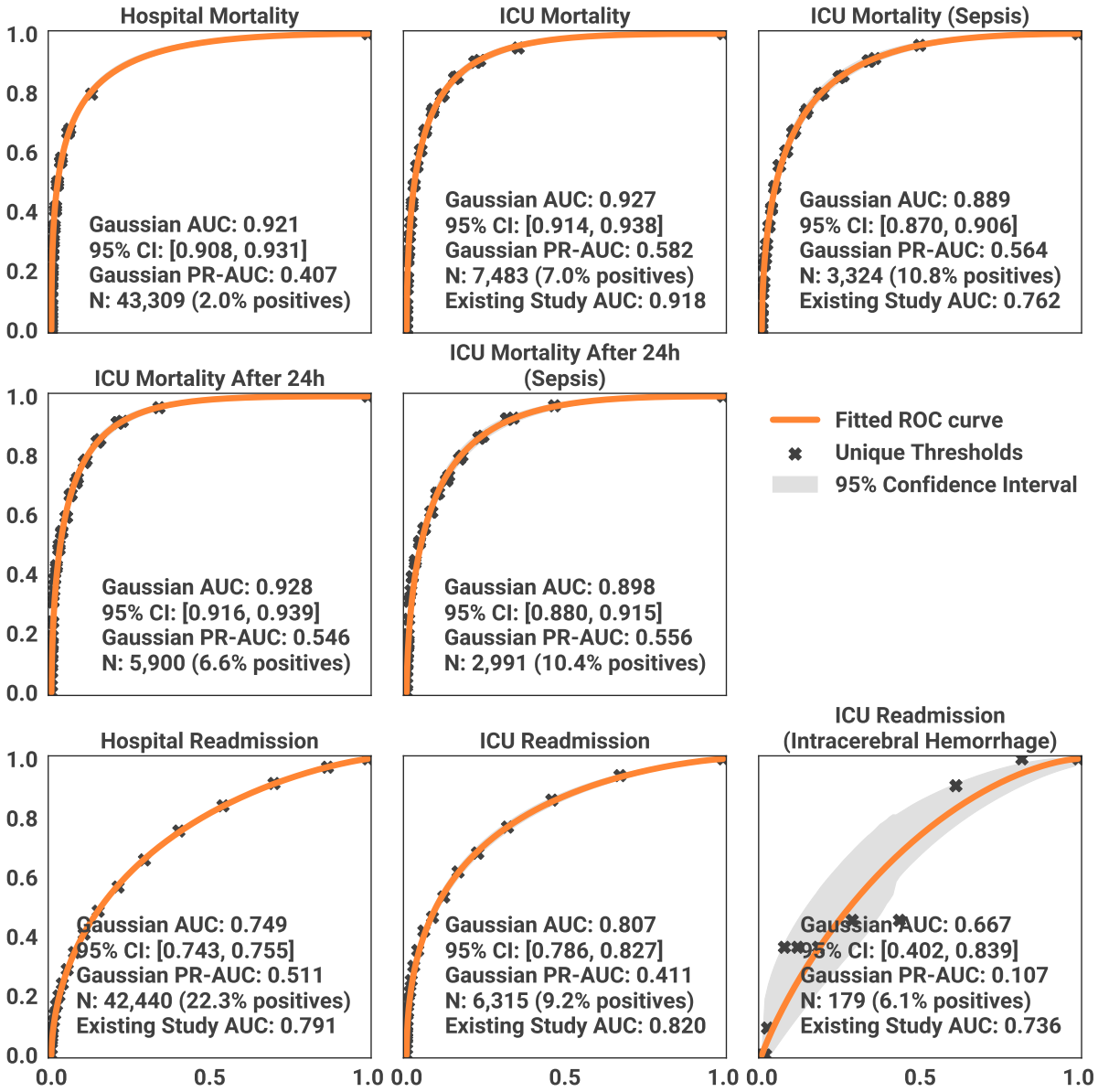


Figure 3.8: **Receiver Operating Characteristic (ROC) curves for mortality and readmission.** ROC curves are shown for inpatient mortality, ICU mortality, and hospital readmission, with shaded regions denoting 95% confidence intervals. Crosses mark selected operating thresholds. ETHOS achieved high discrimination across all outcomes.

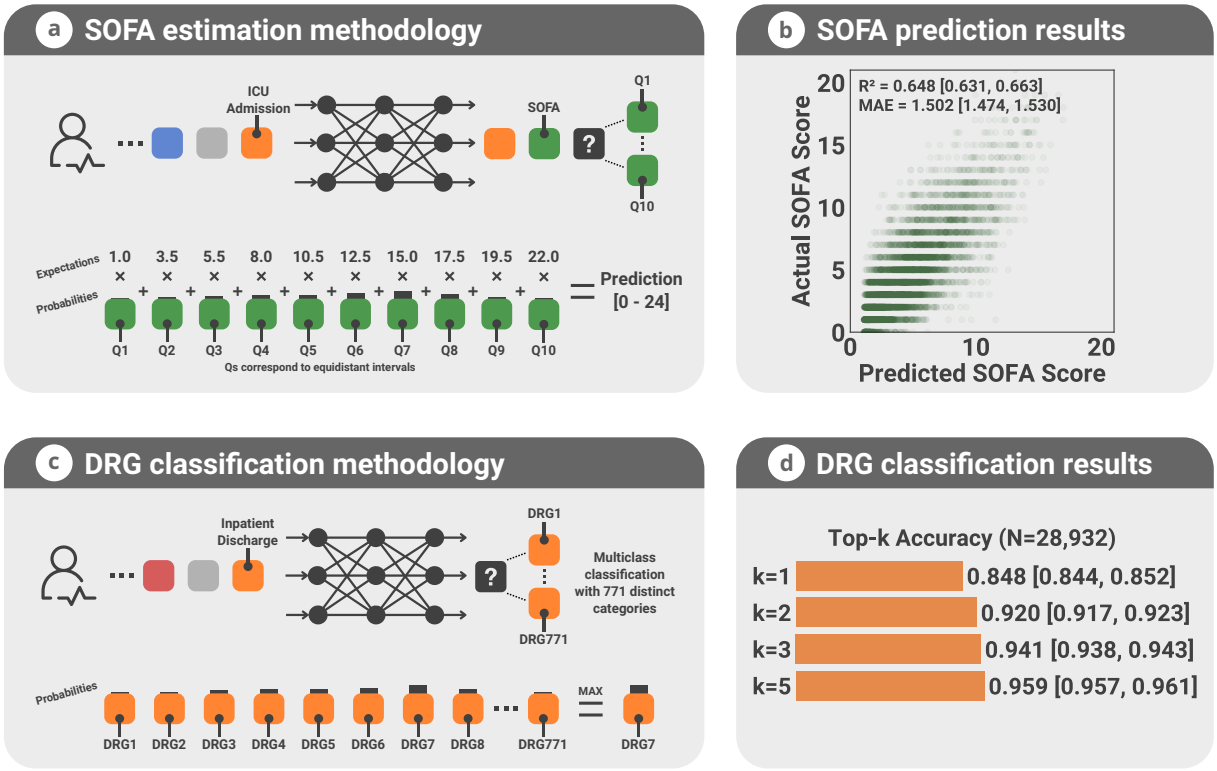
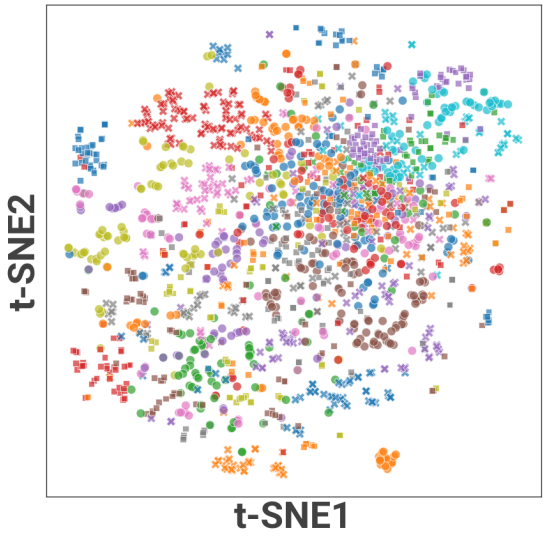
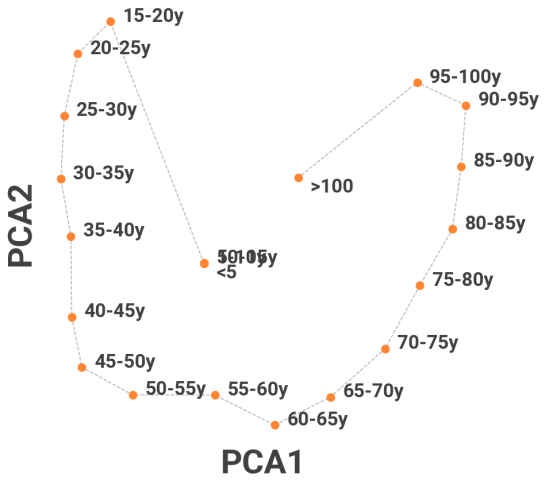


Figure 3.9: **Regression and classification tasks with ETHOS.** (a) Illustration of SOFA token placement. (b) Correlation between predicted and observed SOFA scores. (c) Placement of DRG tokens within patient timelines. (d) DRG classification performance across top-k predictions. These results demonstrate ETHOS's versatility across regression and multiclass classification tasks.



(a) t-SNE projection of embeddings for ICD-10-CM tokens. Colors and shapes denote the first letter of the code (25 diagnostic areas). While not perfectly separable, embeddings reflect clinically relevant clustering by diagnostic category.



(b) PCA projection of quantile tokens encoding patient age and PHT start year. Tokens are arranged in the correct ordinal sequence, showing that ETHOS learned clinically meaningful structure.

Figure 3.10: **Learned token embeddings projected into two dimensions.** Embedding spaces illustrate that ETHOS internalizes semantic and temporal structures without explicit supervision, supporting interpretability.

4. Adaptive Risk Estimation System (ARES)

This chapter is based on the study published in *GigaScience* under the title *Foundation Model of Electronic Medical Records for Adaptive Risk Estimation*. It develops the Enhanced Transformer for Health Outcome Simulation (ETHOS) [40] into the Adaptive Risk Estimation System (ARES), moving from proof-of-concept zero-shot forecasting to a clinically actionable framework for dynamic risk assessment. Whereas ETHOS established that tokenized Patient Health Timelines (PHTs) can be modeled autoregressively to simulate plausible futures without task-specific training, ARES turns this generative capability into continuously updated, patient-specific probability estimates for critical outcomes. In doing so, ARES couples real-time inference with personalized explainability and situates the foundation model within emergency and inpatient workflows, where timely, scenario-aware decision support is essential.

4.1. From Generative Trajectories to Adaptive Risk

The central design principle of ARES is to interpret ETHOS's simulated futures as an empirical distribution over clinically meaningful outcomes. ETHOS, trained on large corpora of tokenized PHTs, can stochastically generate multiple future PHTs (fPHTs) conditioned on a patient's history. ARES aggregates these futures to produce calibrated probabilities for events such as hospital admission from triage, transfer to the intensive care unit, prolonged hospitalization, or in-hospital mortality. This process is conceptually analogous to Monte Carlo simulation; repeated sampling of fPHTs yields an empirical estimate of risk and a natural characterization of uncertainty.

Figure 4.1 summarizes the end-to-end workflow. First, structured electronic health record data (diagnoses, procedures, medications, laboratory values, vital signs, and time intervals) are transformed into a unified token vocabulary and concatenated into a chronological PHT. Second, ETHOS autoregressively samples multiple fPHTs, each a plausible continuation of the patient's trajectory. Third, ARES aggregates these simulated futures to obtain outcome-specific probability distributions and expected timings. Finally, an explainability layer highlights which tokens in the observed PHT are most responsible for the current risk elevation, providing case-level rationale that complements the numeric probabilities.

A defining property of ARES is temporal adaptivity. Whenever new information arrives—an updated vital sign, a new laboratory result, a newly assigned diagnosis—the PHT is extended and the set of fPHTs is regenerated. Estimated risks, therefore, evolve with the patient's clinical course rather than remaining fixed at a single point in time. Figure 4.2 illustrates this behavior in an inpatient course: the risk of intensive care admission rises ahead of transfer; then that component is deactivated once transfer occurs; afterward, the system shifts its attention to prolonged length of stay and discharge disposition, updating probabilities as the trajectory unfolds.

4.2. Inference with Future Patient Health Timelines

Inference in ARES builds directly upon the generative mechanism introduced by ETHOS but formalizes it as a probabilistic estimator that is suitable for bedside decision support. Given a patient's observed PHT prefix, ETHOS samples multiple future PHTs under causal masking until a task-specific stopping criterion is met (for example, discharge, intensive care unit transfer, death, or a prespecified time horizon). Denote these N sampled futures by $\{\tilde{\mathbf{x}}^{(n)}\}_{n=1}^{N}$. For a binary endpoint $\mathcal{E}$ (such as intensive care transfer within a horizon), ARES defines the probability estimate as

$$\widehat{P}(\mathcal{E} \mid \mathbf{x}_{1:L}) \;=\; \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{\mathcal{E} \text{ occurs in } \tilde{\mathbf{x}}^{(n)}\}.$$

For multi-class outcomes (for example, Diagnosis-Related Group at discharge), ARES takes normalized class frequencies across the sampled futures. For regression targets (for example, first-day Sequential Organ Failure Assessment score), ARES extracts the target value from each future (for example, via quantile-to-mean mapping) and averages across samples. In all cases, the distribution of simulated futures supplies not only a point estimate but also a direct measure of dispersion, enabling uncertainty-aware communication with clinicians.

A practical advantage of this scheme is that *competing risks* are handled naturally by the trajectory semantics. If a sampled future terminates in death prior to discharge, then that sample contributes zero probability to prolonged hospitalization; if an early transfer to intensive care occurs, the subsequent risk of intensive care admission is structurally zero in that sample. Because ETHOS generates entire trajectories rather than endpoint labels, the joint and conditional structure among outcomes is preserved by construction. This contrasts with independent classifiers, which can yield incompatible marginal probabilities for causally ordered events.

Finally, ARES is explicitly designed for continuous updating. As the PHT grows, the conditional distribution over plausible futures is resampled, and risk estimates are recomputed. This enables minute-by-minute adaptation in high-acuity settings, where the value of a risk estimate depends on its timeliness as much as its discrimination.

4.3. Methods: From ETHOS (Zero-Shot) to ARES (Adaptive)

The methodological trajectory from ETHOS to ARES reflects a shift from a general demonstration of zero-shot forecasting to a system engineered for clinical deployment: standardized tokenization, scaled training, a formal Monte Carlo inference layer, and patient-level explainability. Below, each axis of change is detailed and contrasted with the original zero-shot study.

4.3.1. Data Processing, Tokenization, and Timeline Assembly

ETHOS (baseline). ETHOS introduced PHTs as a unified representation of heterogeneous events, encoding: (1) diagnoses as ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification), (2) procedures as ICD-10-PCS (International Classification of Diseases, Tenth Revision, Procedure Coding System), (3) medications as ATC (Anatomical Therapeutic Chemical classification), (4) laboratory measurements and vital signs via quantile tokens, (5) admissions and discharges, (6) time-interval tokens (13 bins from minutes to months), and (7) static attributes (age band, sex, race, marital status, and body mass index) reserved in the leading positions of the context window. Events could yield one to seven tokens depending on complexity, and interval tokens were inserted between temporally separated events, after which absolute timestamps were dropped.

ARES (standardized and extended). ARES retained the core vocabulary and causal construction but introduced several standardizations and extensions motivated by deployment needs:

– *Standardized preprocessing*: a MEDS–DEV [1] data pipeline converted raw electronic health records into a stable, auditable intermediate format before tokenization, improving reproducibility and portability across sites (for example, between emergency and inpatient settings).

– *Expanded temporal resolution*: time-interval bins were increased (for example, from thirteen to nineteen) to better capture short gaps in emergency workflows while maintaining long-range seasonal and multi-month tokens for chronic care patterns.

– *Consistent quantiles*: laboratory and vital sign quantiles were recomputed on the larger training corpus (for example, MIMIC-IV v2.2 scale) to stabilize the empirical cumulative distributions across demographic subgroups; age was quantile-encoded to support smooth generalization.

– *Hierarchical codes*: multilevel tokens for the International Classification of Diseases and Anatomical Therapeutic Chemical codes were retained but unified under stricter codebook management so that rare suffixes fall back gracefully to parent-level tokens when data are sparse.

Together, these changes increased representational granularity, reduced site-specific brittleness, and created a consistent token space for emergency department and inpatient tasks.

4.3.2. Model Architecture and Training Configuration

ETHOS (reference training). The original ETHOS models used a decoder-only transformer (GPT-style) with learnable positional embeddings and a context length of 2,048 tokens. Training was performed on concatenated PHTs with next-token prediction, using held-out patients for testing. Hyperparameters were tuned heuristically with a focus on demonstrating zero-shot feasibility across multiple tasks without task-specific finetuning.

ARES (scaled, regularized, and selection-stable). ARES preserved the decoder-only backbone but adopted an explicit configuration after a structured exploration of depth, width, and regularization. The final settings included: 6 transformer blocks, an embedding dimension of 768, 12 attention heads, a context length of 2,048, a dropout of 0.3 throughout the attention and feed-forward layers, the AdamW optimizer with cosine decay (for example, an initial learning rate in the order of 6×10^{-4} decaying to 10^{-5}), and an effective batch size chosen to balance sequence length and device memory. Selection used the running average of recent validation losses to avoid spurious minima. The training scale increased substantially (hundreds of millions of tokens), which improved calibration and subgroup stability. Figure 4.3 summarizes the architecture and hyperparameters used in the final model.

4.3.3. Probabilistic Inference and Calibration

ETHOS (feasibility-level sampling). In the zero-shot study, probabilities were estimated from a modest number of samples per case (for example, twenty futures), sufficient to demonstrate discrimination and to enable the construction of receiver operating characteristic curves across several tasks.

ARES (Monte Carlo formalization). ARES formalized inference as a Monte Carlo estimator with larger sample sizes per patient (for example, at least one hundred futures per endpoint and time horizon), reducing variance and improving probability calibration. Sampling temperature for inference was selected in validation cohorts to balance discrimination and calibration; sampling temperature for synthetic generation (used in ablation and augmentation experiments) was tuned separately to manage the utility–diversity trade-off in generated timelines. In addition to discrimination metrics, ARES tracked calibration with reliability curves and Brier scores, and reported uncertainty

in the form of confidence intervals over bootstrapped cohorts. Ambiguous or degenerate continuations (for example, rare out-of-context emissions) were detected and discarded at a rate below one percent, with thresholds set a priori.

4.3.4. Explainability and Patient-Level Rationale

A major addition in ARES is a token-level attribution mechanism that pairs each updated probability with a ranked list of PHT tokens that are most responsible for the change. Concretely, ARES computes the sensitivity of the outcome probability to perturbations of individual tokens (for example, token removal or embedding masking under a fixed random seed) and aggregates these effects over near-neighbor alternatives. Explanations are displayed as ordered highlights on the PHT with brief, human-readable descriptors (for example, "recent hypoxemia," "broad-spectrum antibiotics started," "new diagnosis: pneumonia"). This provides clinicians not only with a risk number but also with an auditable narrative of why the risk is rising or falling.

4.3.5. Evaluation Design and Benchmarking Suite

Whereas ETHOS validated zero-shot generality on intensive care unit mortality, length of stay, readmission, Sequential Organ Failure Assessment, and Diagnosis-Related Group tasks, ARES targeted decision points encountered at triage and during acute admission. Three benchmark tasks anchored the emergency department evaluation: (1) hospital admission at triage, (2) critical outcome at triage (death or transfer to intensive care within twelve hours), and (3) re-presentation to the emergency department within seventy-two hours after discharge. Additional inpatient tasks included mortality, intensive care transfer, and prolonged length of stay. To contextualize results, ARES was compared with:

- *Clinical scores*: Modified Early Warning Score, National Early Warning Score (versions 1 and 2), Rapid Emergency Medicine Score, Cardiac Arrest Risk Triage, and the Emergency Severity Index [44, 42, 54, 63, 32, 12, 13].

- *Auto-generated scores*: AutoScore for interpretable point systems learned from data [57].

- *Classical machine learning*: Logistic regression, multi-layer perceptron, random forest, and gradient boosting.

- *Deep learning*: Med2Vec for code embeddings and recurrent networks (Long Short-Term Memory) for sequence modeling [8, 15].

- *Strong tabular baseline*: MEDS-Tab for emergency department tabular data (for results context) [33].

All baselines were implemented or adapted following published protocols (with publicly available code for reproducibility), harmonized to common splits and feature availability to ensure a fair comparison.

4.4. Benchmarking Baselines

To make the benchmarking suite self-contained within this chapter, Table 4.1 summarizes the inputs, strengths, and limitations of each comparator method. This table is intended as a quick reference to the methodological continuum against which ARES was evaluated—from static rule-based scores through automatically generated point systems to classical and deep learning models.

Table 4.1: **Summary of benchmarking baselines used for ARES evaluation.** Methods range from hand-crafted clinical scores to modern deep learning sequence models. Inputs, advantages, and limitations are listed relative to the demands of longitudinal, heterogeneous electronic health record modeling.

Category	Model	Key Inputs	Strengths / Limitations
Clinical scores	MEWS [44]	Vital signs; level of consciousness	Simple and ubiquitous; static thresholds; no temporal modeling.
	NEWS v1/v2 [42, 54, 63]	Vital signs; oxygenation; supp. oxygen	Embedded in workflows; still snapshot-based; not patient-specific.
	REMS [32]	Vital signs; age	Emergency department focus; additive; static by design.
	CART [12]	Vital sign features in logistic regression	Greater statistical rigor; limited features; no trajectory context.
	ESI [13]	Nurse-assigned five-level triage	Interpretable and standard; subjective; not data-adaptive.
Auto-generated	AutoScore [57]	Ranked and binned EHR features	Transparent point-based scores; cannot capture long-range temporal structure.
Classical ML	Logistic Regression	Fixed-length vectors	Interpretable linear effects; limited to pre-engineered features.
	Multi-layer Perceptron	Fixed-length vectors	Models nonlinearity; ignores sequence order and timing.
	Random Forest	Tabular features	Robust ensembles; static snapshots; reduced interpretability.
	Gradient Boosting	Tabular features	Strong tabular performance; careful tuning needed; no temporal dynamics.
Deep learning	Med2Vec [8]	Codes across visits	Learns distributed code embeddings; weak temporal expressivity.
	Long Short-Term Memory [15]	Visit sequences or event streams	Captures short- to mid-range temporal patterns; struggles with very long, irregular timelines.

4.5. Datasets

ARES builds upon the ETHOS foundation but extends its evaluation to cover a broader spectrum of clinical encounters. The primary source of training and evaluation data was the MIMIC-IV v2.2 database, which provides longitudinal inpatient and ICU records for over 200,000 patients. MIMIC-IV encompasses detailed information on admissions, diagnoses, procedures, laboratory results, medications, and outcomes, making it a standard benchmark for predictive modeling in critical care.

To further stress-test model generalization, we augmented this dataset with the **MIMIC-ED** cohort. MIMIC-ED contains emergency department (ED) encounters, capturing a distinct phase of patient care characterized by shorter stays, rapid decision-making, and heterogeneous presenting complaints. Unlike inpatient admissions, which typically include extended monitoring and complete diagnostic workups, ED encounters provide high-temporal-resolution snapshots of early disease presentation and stabilization efforts. The inclusion of MIMIC-ED therefore added two critical dimensions to model evaluation: (i) the ability to process incomplete or rapidly evolving records at the point of first contact, and (ii) the opportunity to assess whether the predictive performance observed in structured inpatient data could transfer to more acute and fragmented ED settings. By unifying MIMIC-IV and MIMIC-ED through the standardized Patient Health Timeline (PHT) tokenization scheme, ARES was evaluated under conditions that span the emergency-to-inpatient continuum of care.

4.6. Synthesis: What Changes from ETHOS to ARES

In aggregate, ARES preserves ETHOS's generality while introducing architectural and engineering advances necessary for clinical viability:

1. **Representation.** The PHT token space is standardized and expanded, including finer-grained temporal intervals, harmonized quantile encodings, and formal codebook governance. This provides a stable interface across emergency and inpatient records.

2. **Scale.** Training is scaled in terms of tokens, patients, and iterations. Regularization protocols and model selection strategies improve subgroup stability and probability calibration.

3. **Inference.** ARES formalizes generative sampling with a Monte Carlo inference layer. By sampling large numbers of future trajectories and explicitly handling ambiguous continuations, ARES aligns discrimination with reliability while maintaining causal consistency.

4. **Explainability.** Case-level rationales are generated alongside every risk update, highlighting the tokens most influential to the prediction. This improves interpretability and facilitates bedside adoption.

5. **Benchmarking.** ARES is compared against strong baselines under harmonized conditions. Improvements are demonstrated not only in aggregate accuracy but also in subgroup robustness and calibration, underscoring the qualitative benefits of trajectory-aware, zero-shot inference.

The following sections apply these methods to inpatient and ED cohorts to quantify discrimination, calibration, and interpretability while comparing adaptive scenario-based risk estimation against snapshot-based tabular alternatives.

4.7. Core Hospital Prediction Tasks

We first evaluated ARES on four clinically consequential endpoints that together summarize major inpatient risks: (1) **Hospital mortality**, defined as death during the index hospitalization; (2) **ICU admission**, representing an escalation from the general ward to intensive care; (3) **Prolonged length of stay**, defined as a hospital stay exceeding ten days; and (4) a **composite outcome**, combining mortality, ICU admission, and prolonged stay to reflect the overall burden of deterioration and resource use.

This composite endpoint is especially valuable, as it forces the model to reason about competing risks across the trajectory: once death occurs in a sampled timeline, a prolonged stay cannot follow. Such consistency is not easily achievable with independent task-specific classifiers.

Figure 4.4 presents discrimination results for ARES (orange) compared with MEDS-Tab (gray), the strongest tabular baseline. ARES consistently achieved higher AUC values across all four tasks, both in the aggregate population and within gender and race subgroups. Improvements were not only statistically significant but also clinically meaningful, reflecting fewer missed deteriorations and fewer false alerts at comparable thresholds.

4.7.1. Detailed Results by Subgroup

Table 4.2 reports AUCs with 95% confidence intervals, alongside subgroup-level prevalence estimates. Several important observations arise:

– **Overall performance.** ARES achieved an AUC of 0.940 for hospital mortality, 0.932 for ICU admission, 0.853 for prolonged stay, and 0.906 for the composite endpoint. Each represents a consistent improvement over MEDS-Tab, with the most pronounced gain for mortality (+0.053 absolute AUC). This reflects the advantage of timeline-based simulation in detecting rare but critical outcomes.

– **Gender subgroups.** Performance remained stable across male and female patients, with negligible differences in AUC (within ± 0.01). This indicates that ARES generalizes effectively across genders, an important property given prior reports of bias in predictive models trained on skewed datasets.

– **Race subgroups.** Across racial categories, ARES consistently outperformed MEDS-Tab; however, the magnitude varied. Particularly strong gains were observed among Asian, Black, and Hispanic patients, with mortality AUCs exceeding 0.95 in some groups. These improvements suggest that ETHOS's standardized tokenization, coupled with ARES's scaled training, mitigates disparities that can arise when models rely heavily on snapshot-level coding systems.

– **Other and Unknown categories.** ARES demonstrated very high AUCs for the "Other" group (0.985 for mortality, 0.956 for ICU transfer), though confidence intervals were wider due to smaller sample sizes. By contrast, the "Unknown" category exhibited lower mortality discrimination (0.886) despite high AUCs for ICU transfer and the composite endpoint. This highlights the challenges posed by missing demographic data, underscoring the need for careful calibration when demographic covariates are incomplete.

– **Composite outcome.** ARES achieved consistent advantages on the composite task across all subgroups. By leveraging fully simulated timelines, the model ensures coherent risk profiles: once a terminal event (death) is generated, subsequent outcomes (e.g., prolonged stay) are excluded, preserving causal logic. This provides clinicians with a more reliable risk estimate compared to independent classifiers, which may produce logically inconsistent predictions.

4.7.2. Emergency Department Benchmarking

To assess performance at the earliest point of contact, ARES was evaluated on emergency department (ED) prediction tasks, where timeliness and reliability are paramount. The triage setting provides only a sparse snapshot of the patient's state; consequently, many existing methods either rely on fixed early warning scores or tabular classifiers built from a limited feature set. By contrast, ARES consumes the full Patient Health Timeline accumulated up to triage and stochastically generates plausible futures conditioned on that history.

At triage, ARES achieved strong discrimination for **hospitalization at triage**, with an area under the ROC curve of 0.912, surpassing all rule-based scores (for example, Modified Early Warning Score, National Early Warning Score, Cardiac Arrest Risk Triage) and machine learning baselines (for example, gradient boosting, multi-layer perceptron). For the **critical outcome within 12 hours**, defined as intensive care unit admission or death within twelve hours of triage, ARES reached an area under the ROC curve of 0.937, indicating excellent early identification of patients likely to deteriorate rapidly. This task benefits markedly from token-level representations of acute physiology (for example, high-quantile lactate, hypotension bins) and from explicit time-interval tokens that capture the tempo of clinical change rather than only its magnitude.

For **re-presentation within 72 hours of ED discharge**, a notoriously multifactorial endpoint influenced by clinical and social determinants, ARES obtained an area under the ROC curve of 0.740. Although lower than the acute deterioration tasks, this performance exceeded the best competing baselines, suggesting that timeline-aware generation confers an advantage even for outcomes with complex, nonclinical drivers. Here, embedding patterns over prior utilization (for example, repeated short-interval admissions, recurrent diagnoses) contribute substantially to discrimination.

4.7.3. Summary of Benchmarking

Taken together, the benchmarking results show that ARES provides consistently superior discrimination across inpatient and emergency department settings while maintaining strong calibration and subgroup robustness. Figure 4.6 summarizes receiver operating characteristic (ROC) curves for all seven evaluated tasks. In the inpatient cohort, hospital mortality reached an AUROC of 0.940 (95% CI: 0.932–0.946), ICU transfer 0.932 (95% CI: 0.928–0.935), prolonged length of stay 0.853 (95% CI: 0.848–0.858), and the composite endpoint 0.906 (95% CI: 0.902–0.909). In the emergency department cohort, hospitalization at triage achieved 0.946 (95% CI: 0.944–0.947), the twelve hour critical outcome reached 0.945 (95% CI: 0.940–0.950), and re presentation within seventy two hours achieved 0.745 (95% CI: 0.729–0.763). These results indicate that the same timeline based generative machinery extends naturally from inpatient settings to the faster paced and more fragmented ED environment.

We next detail ED task performance, referencing Tables 4.3–4.5. Across all three ED endpoints, ARES delivered large absolute gains over classical scores and tabular or neural baselines in both ranking metrics and operating point performance.

Hospitalization at triage. Table 4.3 reports AUROC, AUPRC, and thresholded sensitivity and specificity. Traditional triage scores, including ESI, NEWS, NEWS2, REMS, MEWS, and CART, underperformed learning based baselines, reflecting limited capacity to integrate multimodal history at arrival. Among machine learning comparators, MEDS Tab was the strongest, with AUROC 0.863 and AUPRC 0.879. ARES substantially improved upon this benchmark, reaching AUROC 0.946 and AUPRC 0.945, and demonstrating balanced operating characteristics at the ROC optimal point (sensitivity 0.868, specificity 0.864). The improvement is clinically meaningful because it simultaneously raises true positive capture at arrival and reduces false admission recommendations, a desirable profile for crowding constrained ED workflows.

Twelve hour critical outcome. Table 4.4 shows that ARES improves event ranking and early warning compared with triage scores and machine learning baselines. MEDS Tab achieved AUROC 0.853 and AUPRC 0.513, whereas ARES reached AUROC 0.945 and AUPRC 0.696. At the ROC optimal point, sensitivity and specificity were both high (0.876 and 0.873), which supports early escalation decisions without excessive alarm burden. Gains are largest in AUPRC, consistent with improved precision for low prevalence critical events.

ED re presentation within seventy two hours. Table 4.5 shows that representation is the most challenging ED endpoint, with a lower absolute AUPRC across all methods due to low event prevalence and heterogeneous causes. Even so, ARES outperformed all comparators, with AUROC 0.745 and AUPRC 0.214, improving upon MEDS Tab (AUROC 0.714, AUPRC 0.189). At the ROC optimal point, sensitivity and specificity were balanced (0.669 and 0.685), which helps identify likely return visits without overwhelming clinicians with false positives. The gains suggest that timeline aware features, for example, recent utilization patterns and evolving symptom codes, add predictive value beyond static demographics and index visit vitals.

Calibration. Calibration analysis in Figure 4.7 confirms that ARES's probabilistic outputs align with observed event frequencies across tasks. Brier scores were excellent for hospital mortality (0.015), the twelve hour critical outcome (0.030), and ED re presentation within seventy two hours (0.041). Predictions for ICU transfer (0.062), prolonged stay (0.067), and the composite endpoint (0.088) tracked the identity line with minimal deviation. Hospitalization at triage displayed acceptable calibration (0.094), with mild overestimation at the top decile of predicted risk, a common pattern when decision thresholds prioritize sensitivity at arrival. These findings indicate that Monte Carlo aggregation over simulated futures yields not only accurate rankings but also well calibrated probabilities suitable for thresholding, triage, and shared decision making.

4.8. Explainability and Personalized Risk Factors

A central challenge for predictive models in healthcare is interpretability, as clinicians must be able to understand not only the quantitative risk estimates but also the specific factors driving those predictions. ARES addresses this by augmenting probabilistic outputs with token-level attributions derived directly from simulated futures. These attributions quantify the marginal contribution of individual events in the Patient Health Timeline (PHT) to shifts in predicted risk. For example, when a token such as a high-quantile laboratory measurement or a procedure code consistently changes the outcome distribution across generated trajectories, its effect is surfaced as a personalized explanation.

Figure 4.8 illustrates this mechanism. The lower panel displays evolving probabilities for hospital mortality, ICU admission, prolonged length of stay, and the composite endpoint as new data are sequentially added. The upper panel highlights regions where specific tokens exert a strong marginal influence on predicted risks. For instance, a sharp rise in ICU transfer probability coincides with the token corresponding to endotracheal intubation (ICD-10-PCS 0BH17EZ), while the subsequent initiation of mechanical ventilation (ICD-10-PCS 5A12012) elevates mortality risk but suppresses further ICU transfer probability, reflecting the causal consistency enforced by full-trajectory simulation. Earlier in the sequence, a high-quantile lactate measurement temporarily increases composite risk, an effect that is attenuated once subsequent normal values are introduced. This pattern mirrors established clinical reasoning, where persistently abnormal values carry greater prognostic weight than transient deviations.

By grounding explanations in a discrete and clinically interpretable vocabulary (ICD-10-CM for diagnoses, ICD-10-PCS for procedures, ATC for drugs, quantile-encoded laboratory and vital signs, and explicit time-interval tokens), ARES ensures that rationales are aligned with familiar medical semantics rather than opaque latent features. This transparency supports a range of clinical use cases: early warning with justification, prioritization of scarce monitoring resources, and counterfactual reasoning (for example, simulating how risk would change if a procedure were absent). The ability to link outcome probabilities to specific, recognizable events enhances trust and facilitates bedside adoption.

4.9. Discussion and Clinical Relevance

The results demonstrate that ARES operationalizes ETHOS's generative capabilities into a clinically viable adaptive risk estimation system. Across both inpatient and emergency department settings, ARES achieved state-of-the-art discrimination (Figures 4.4, 4.6, 4.7), robust calibration, and consistency across demographic subgroups (Table 4.2). Importantly, the probabilistic outputs are not static but evolve as new events appear in a patient's record, allowing clinicians to track trajectories of risk in real time. This dynamic updating enables earlier recognition of deterioration, targeted escalation for high-risk patients, and reassurance when trajectories stabilize, supporting both patient safety and resource stewardship.

Uncertainty quantification is another distinctive advantage. By sampling multiple future PHTs, ARES provides not only a point estimate of risk but also confidence intervals and trajectory variability. This allows clinicians to assess not only whether a patient is at high risk but also how stable that risk estimate is under plausible futures, a feature that is absent from most traditional early warning systems.

Equally important is the equity of performance. Stratified analyzes show that ARES maintains high discrimination across gender and racial subgroups, with particularly strong gains compared to MEDS-Tab in Asian, Black, and Hispanic patients (Figure 4.4, Table 4.2). While continued fairness auditing remains essential, these results suggest that leveraging longitudinal context may mitigate biases that arise in snapshot-based models trained on narrower sets of features.

From a systems perspective, the reliance on a discrete, interoperable token vocabulary provides additional benefits. The same tokenization that enables interpretability also facilitates cross-institutional deployment and federated training, as institutions can align data to a common representation without sharing raw records. This portability makes ARES suitable not only for single-site adoption but also for collaborative, privacy-preserving model development at scale.

Finally, ARES provides a blueprint for extending ETHOS beyond its initial scope. As demonstrated in Section 4.7.3, the same framework can accommodate a wide spectrum of outcomes, from short-term deterioration (twelve-hour critical outcome at triage) to longer-term resource utilization (prolonged stay). Additional metrics—such as acute kidney injury, sepsis, or readmission risk—can be layered on without retraining, since they can be computed from the same set of simulated trajectories. This extensibility highlights the promise of timeline-native foundation models as general-purpose engines for clinical prediction.

In conclusion, ARES transforms the theoretical advances of ETHOS into a practical, adaptive, and interpretable system for real-world healthcare. By uniting generative trajectory modeling with calibrated probabilities and token-level explanations, it advances the field toward foundation models that are not only accurate but also trustworthy, equitable, and deployable across diverse clinical environments.

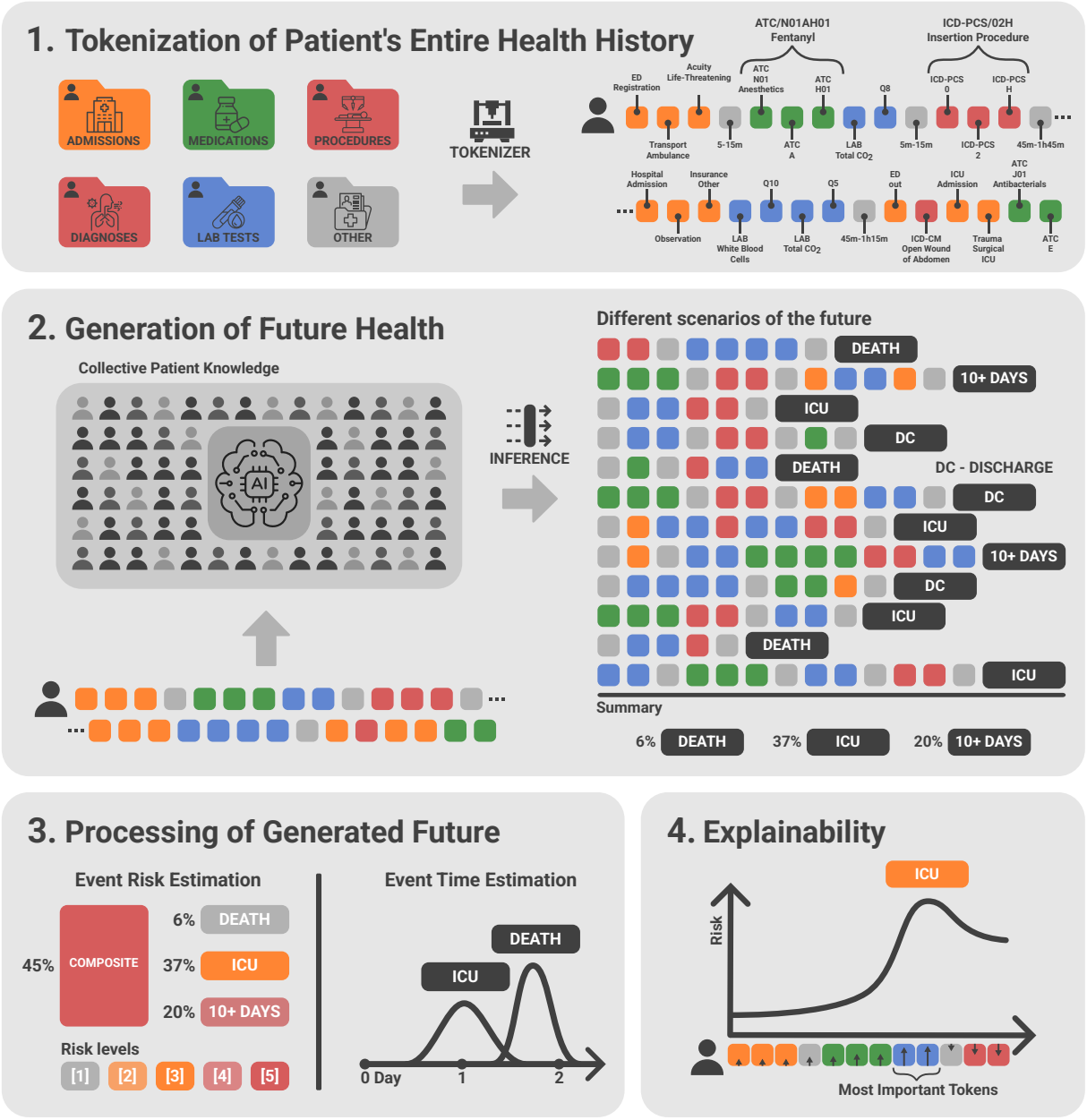


Figure 4.1: **ARES pipeline.** (*Left*) Tokenization and assembly of the Patient Health Timeline (PHT) from heterogeneous electronic health record sources. (*Middle*) Autoregressive sampling of multiple future PHTs (fPHTs) by the ETHOS generator. (*Right*) Aggregation of futures into calibrated, time-aware risk estimates with token-level explainability.

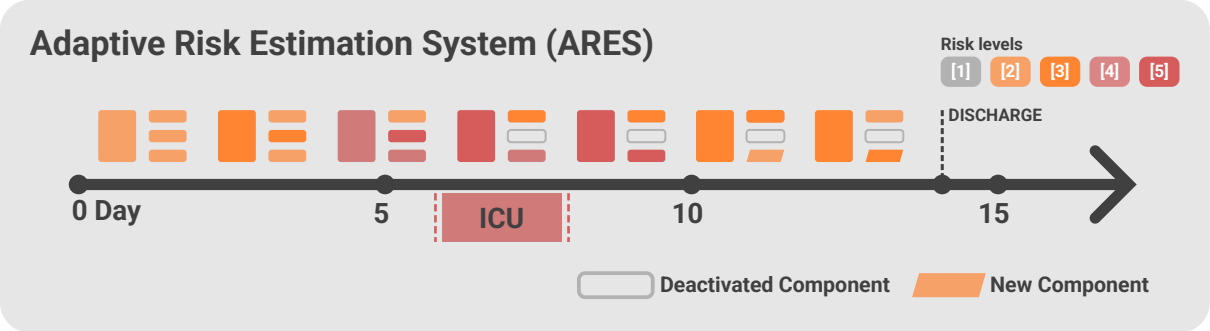
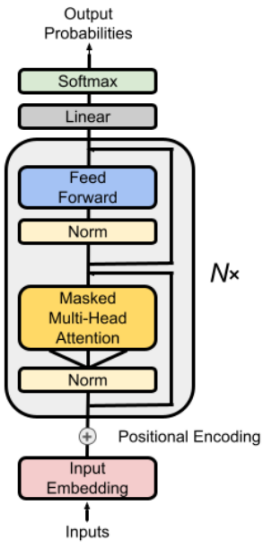


Figure 4.2: **Dynamic risk along a hospital course.** ARES updates probabilities as the PHT accrues new events. Resolved components (e.g., intensive care admission) are deactivated; emerging concerns (e.g., prolonged length of stay) are activated and tracked with expected timing.



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 4.3: **ARES architecture and hyperparameters.** The final model uses a decoder-only transformer with 6 blocks, embedding dimension 768, 12 attention heads, context length 2,048, and dropout 0.3. Optimization uses AdamW with a decaying learning rate; model selection averages recent validation losses.

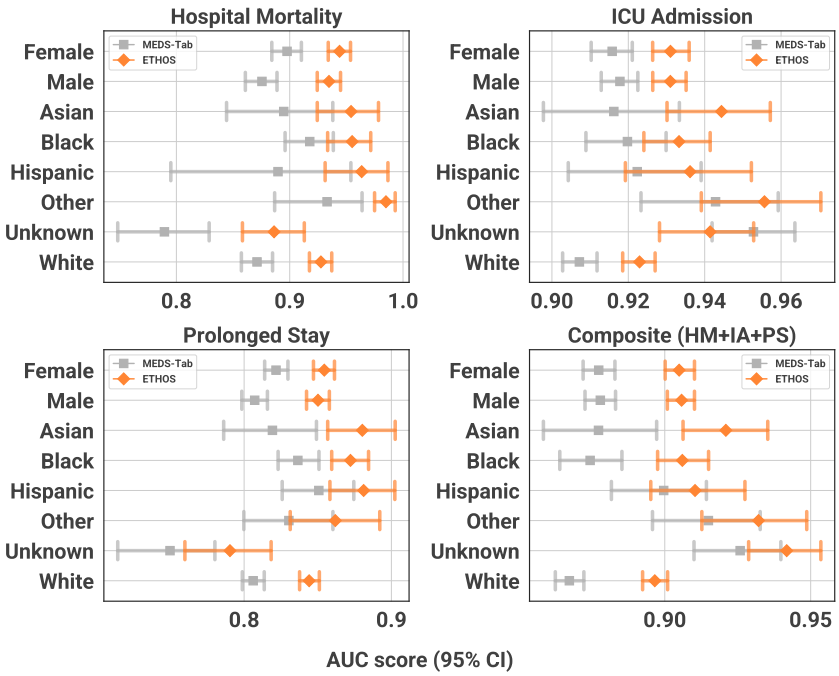


Figure 4.4: **Discrimination across core hospital outcomes.** ARES (orange) versus MEDS-Tab (gray) for hospital mortality, ICU admission, prolonged stay, and the composite outcome. Results are shown overall and stratified by gender and race. Vertical bars denote 95% confidence intervals.

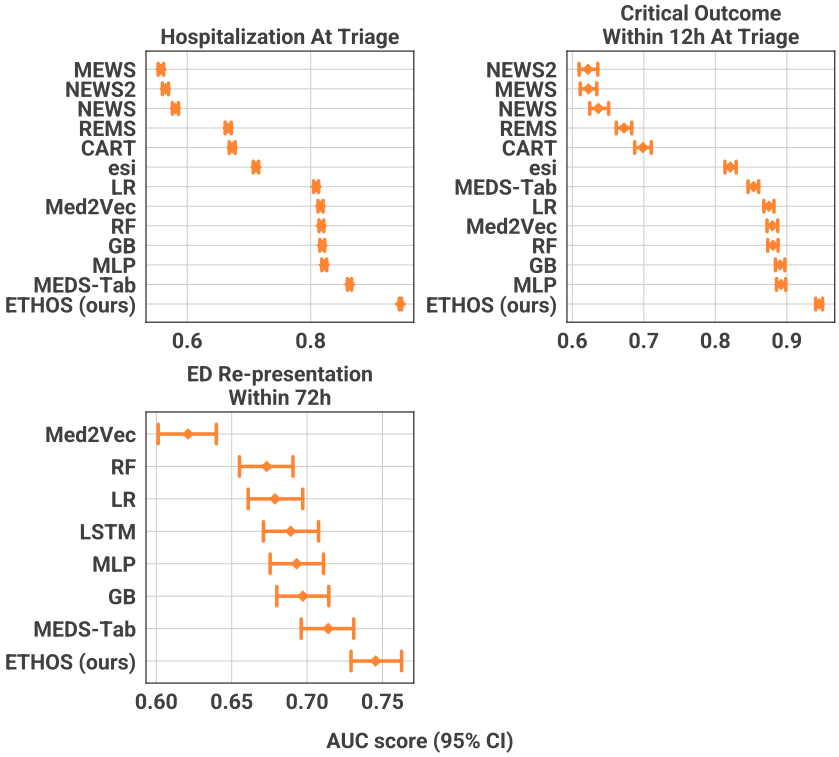


Figure 4.5: **Emergency department tasks at triage.** ARES performance for hospitalization at triage, critical outcome within twelve hours (ICU transfer or death), and re-presentation within seventy-two hours after discharge. Bars show area under the ROC curve with 95% confidence intervals; dashed lines indicate performance of representative baselines.

Table 4.2: Area under the ROC curve (AUC) with 95% confidence intervals by subgroup (gender and race) for hospital mortality, ICU admission, prolonged stay, and the composite outcome. Prevalence values are reported for the full cohort.

	Hospital Mortality	**ICU Admission**	**Prolonged Stay**	**Composite (HM+IA+PS)**
Prevalence (%)	1.95	15.44	9.01	20.41
		ARES		
Overall	0.940 [0.932, 0.947]	0.932 [0.928, 0.935]	0.853 [0.848, 0.858]	0.906 [0.902, 0.909]
Gender				
Female	0.944 [0.933, 0.953]	0.931 [0.927, 0.936]	0.854 [0.847, 0.862]	0.905 [0.900, 0.910]
Male	0.935 [0.924, 0.945]	0.931 [0.926, 0.936]	0.850 [0.842, 0.857]	0.906 [0.901, 0.911]
Race				
Asian	0.954 [0.925, 0.978]	0.944 [0.930, 0.958]	0.880 [0.858, 0.900]	0.921 [0.905, 0.936]
Black	0.955 [0.935, 0.972]	0.933 [0.924, 0.943]	0.872 [0.858, 0.885]	0.906 [0.897, 0.915]
Hispanic	0.964 [0.929, 0.987]	0.936 [0.918, 0.951]	0.881 [0.856, 0.903]	0.910 [0.893, 0.927]
Other	0.985 [0.975, 0.993]	0.956 [0.937, 0.972]	0.862 [0.831, 0.889]	0.932 [0.915, 0.949]
Unknown	0.886 [0.852, 0.911]	0.941 [0.928, 0.953]	0.790 [0.757, 0.819]	0.942 [0.928, 0.954]
White	0.928 [0.918, 0.937]	0.923 [0.919, 0.927]	0.844 [0.838, 0.851]	0.897 [0.892, 0.901]
		MEDS-Tab		
Overall	0.887 [0.877, 0.897]	0.918 [0.914, 0.921]	0.815 [0.810, 0.821]	0.879 [0.875, 0.883]
Gender				
Female	0.898 [0.884, 0.910]	0.916 [0.910, 0.921]	0.822 [0.814, 0.830]	0.877 [0.872, 0.883]
Male	0.876 [0.861, 0.889]	0.918 [0.913, 0.922]	0.807 [0.798, 0.816]	0.878 [0.873, 0.883]
Race				
Asian	0.895 [0.844, 0.938]	0.916 [0.898, 0.933]	0.819 [0.786, 0.849]	0.877 [0.858, 0.897]
Black	0.918 [0.896, 0.938]	0.920 [0.909, 0.930]	0.836 [0.823, 0.851]	0.874 [0.864, 0.885]
Hispanic	0.890 [0.795, 0.954]	0.922 [0.904, 0.939]	0.851 [0.826, 0.874]	0.900 [0.882, 0.914]
Other	0.933 [0.887, 0.964]	0.943 [0.923, 0.959]	0.830 [0.800, 0.860]	0.915 [0.896, 0.933]
Unknown	0.789 [0.748, 0.829]	0.953 [0.942, 0.964]	0.750 [0.714, 0.780]	0.926 [0.910, 0.940]
White	0.871 [0.857, 0.885]	0.907 [0.903, 0.912]	0.806 [0.799, 0.814]	0.867 [0.862, 0.872]

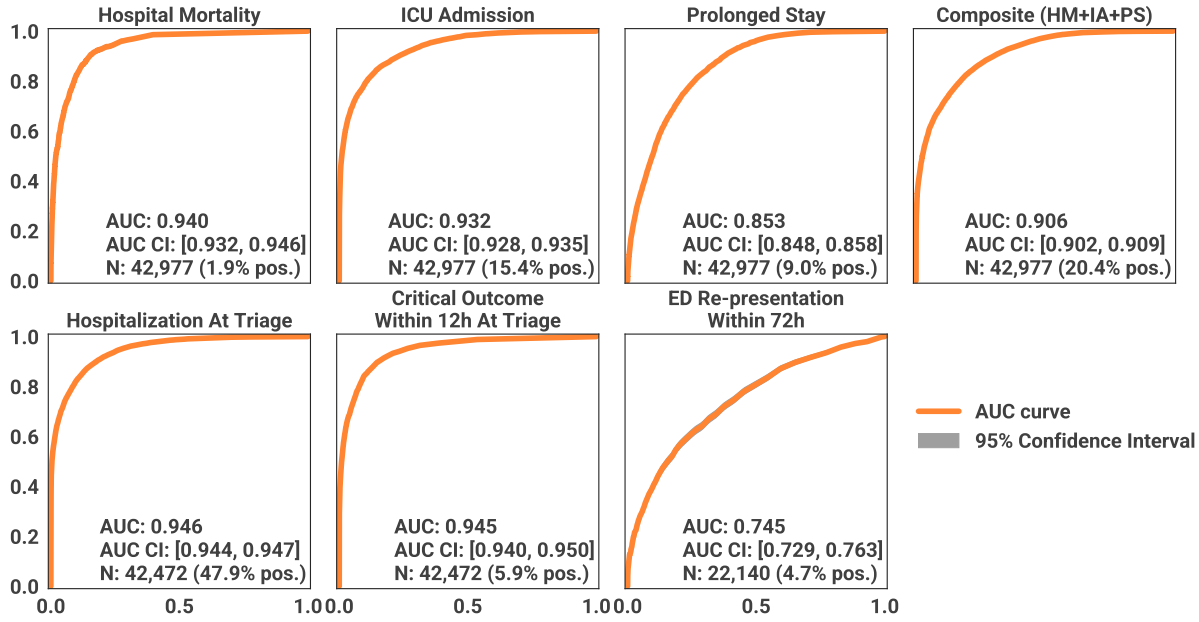


Figure 4.6: **Discrimination across inpatient and emergency department tasks.** Receiver operating characteristic curves for seven endpoints: hospital mortality, ICU transfer, prolonged length of stay, composite inpatient outcome, hospitalization at triage, twelve hour critical outcome, and ED re presentation within seventy two hours. Shaded bands denote 95% confidence intervals estimated by bootstrap.

Table 4.3: **Hospitalization at triage.** Performance comparison for predicting admission disposition at ED triage. Metrics include AUROC, AUPRC, and sensitivity and specificity at the ROC point closest to (0, 1), with 95% confidence intervals in brackets. ARES, implemented with the standardized PHT interface, outperforms triage scores and machine learning baselines, including the strongest tabular comparator, MEDS Tab.

	AUROC	AUPRC	Sensitivity	Specificity
LR	0.809 [0.805, 0.813]	0.775 [0.769, 0.781]	0.734 [0.721, 0.750]	0.736 [0.721, 0.749]
Med2Vec	0.816 [0.812, 0.820]	0.782 [0.775, 0.788]	0.751 [0.734, 0.770]	0.728 [0.711, 0.745]
RF	0.817 [0.814, 0.821]	0.785 [0.779, 0.791]	0.759 [0.736, 0.764]	0.726 [0.720, 0.747]
GB	0.819 [0.815, 0.823]	0.792 [0.786, 0.798]	0.753 [0.731, 0.770]	0.728 [0.715, 0.751]
MLP	0.822 [0.818, 0.826]	0.796 [0.790, 0.802]	0.754 [0.743, 0.775]	0.734 [0.716, 0.745]
ESI	0.712 [0.707, 0.716]	0.632 [0.625, 0.638]	0.584 [0.577, 0.590]	0.784 [0.779, 0.789]
NEWS	0.581 [0.576, 0.586]	0.555 [0.548, 0.561]	0.563 [0.556, 0.569]	0.546 [0.540, 0.553]
NEWS2	0.565 [0.560, 0.570]	0.538 [0.532, 0.544]	0.519 [0.512, 0.526]	0.570 [0.564, 0.577]
REMS	0.666 [0.661, 0.671]	0.605 [0.598, 0.612]	0.605 [0.552, 0.722]	0.641 [0.545, 0.711]
MEWS	0.558 [0.553, 0.562]	0.521 [0.515, 0.527]	0.296 [0.289, 0.302]	0.812 [0.806, 0.817]
CART	0.673 [0.668, 0.678]	0.617 [0.610, 0.624]	0.703 [0.696, 0.709]	0.578 [0.571, 0.585]
MEDS-Tab	0.863 [0.860, 0.866]	0.879 [0.876, 0.883]	0.746 [0.735, 0.754]	0.820 [0.809, 0.835]
ARES (ours)	0.946 [0.944, 0.947]	0.945 [0.943, 0.947]	0.868 [0.859, 0.876]	0.864 [0.856, 0.873]

Table 4.4: **Critical outcome within twelve hours at triage.** Performance comparison for predicting a composite critical event within twelve hours of ED triage. Metrics include AUROC, AUPRC, and sensitivity and specificity at the ROC point closest to (0, 1), with 95% confidence intervals in brackets. ARES achieves the highest performance across all metrics, substantially surpassing triage scores and machine learning baselines.

	AUROC	AUPRC	Sensitivity	Specificity
LR	0.875 [0.868, 0.882]	0.308 [0.288, 0.328]	0.813 [0.792, 0.836]	0.782 [0.766, 0.803]
Med2Vec	0.880 [0.872, 0.887]	0.324 [0.305, 0.346]	0.817 [0.799, 0.852]	0.787 [0.762, 0.804]
RF	0.881 [0.873, 0.888]	0.362 [0.343, 0.386]	0.812 [0.794, 0.829]	0.792 [0.788, 0.796]
GB	0.891 [0.884, 0.897]	0.389 [0.367, 0.412]	0.836 [0.804, 0.848]	0.788 [0.779, 0.812]
MLP	0.892 [0.886, 0.898]	0.372 [0.352, 0.396]	0.845 [0.806, 0.855]	0.784 [0.780, 0.823]
ESI	0.821 [0.814, 0.829]	0.190 [0.178, 0.201]	0.900 [0.887, 0.913]	0.637 [0.632, 0.642]
NEWS	0.637 [0.624, 0.651]	0.139 [0.127, 0.154]	0.461 [0.440, 0.483]	0.796 [0.792, 0.801]
NEWS2	0.622 [0.610, 0.636]	0.130 [0.118, 0.144]	0.416 [0.402, 0.605]	0.822 [0.533, 0.826]
REMS	0.672 [0.661, 0.683]	0.093 [0.087, 0.102]	0.662 [0.642, 0.683]	0.602 [0.597, 0.607]
MEWS	0.623 [0.611, 0.634]	0.101 [0.093, 0.110]	0.445 [0.424, 0.466]	0.772 [0.768, 0.776]
CART	0.699 [0.687, 0.710]	0.134 [0.123, 0.147]	0.579 [0.557, 0.600]	0.720 [0.716, 0.725]
MEDS-Tab	0.853 [0.846, 0.861]	0.513 [0.493, 0.531]	0.735 [0.717, 0.752]	0.764 [0.759, 0.771]
ARES (ours)	0.945 [0.941, 0.950]	0.696 [0.678, 0.712]	0.876 [0.860, 0.898]	0.873 [0.852, 0.889]

Table 4.5: **ED re presentation within seventy two hours.** Performance comparison for predicting return visits within seventy two hours of discharge. Metrics include AUROC, AUPRC, and sensitivity and specificity at the ROC point closest to (0, 1), with 95% confidence intervals in brackets. ARES outperforms triage scores and learning based baselines on this difficult, low prevalence endpoint.

	AUROC	AUPRC	Sensitivity	Specificity
LR	0.679 [0.661, 0.697]	0.161 [0.141, 0.185]	0.562 [0.544, 0.645]	0.699 [0.613, 0.718]
Med2Vec	0.621 [0.601, 0.640]	0.128 [0.110, 0.148]	0.560 [0.477, 0.595]	0.615 [0.568, 0.725]
RF	0.673 [0.655, 0.691]	0.150 [0.131, 0.173]	0.642 [0.549, 0.665]	0.599 [0.594, 0.693]
GB	0.697 [0.680, 0.714]	0.165 [0.143, 0.188]	0.623 [0.592, 0.704]	0.662 [0.582, 0.690]
MLP	0.693 [0.675, 0.711]	0.168 [0.147, 0.192]	0.603 [0.579, 0.676]	0.675 [0.607, 0.701]
LSTM	0.689 [0.671, 0.708]	0.164 [0.143, 0.186]	0.595 [0.566, 0.653]	0.680 [0.633, 0.722]
MEDS-Tab	0.714 [0.696, 0.731]	0.189 [0.167, 0.214]	0.645 [0.575, 0.689]	0.657 [0.617, 0.742]
ARES (ours)	0.745 [0.728, 0.762]	0.214 [0.190, 0.239]	0.669 [0.611, 0.698]	0.685 [0.657, 0.757]

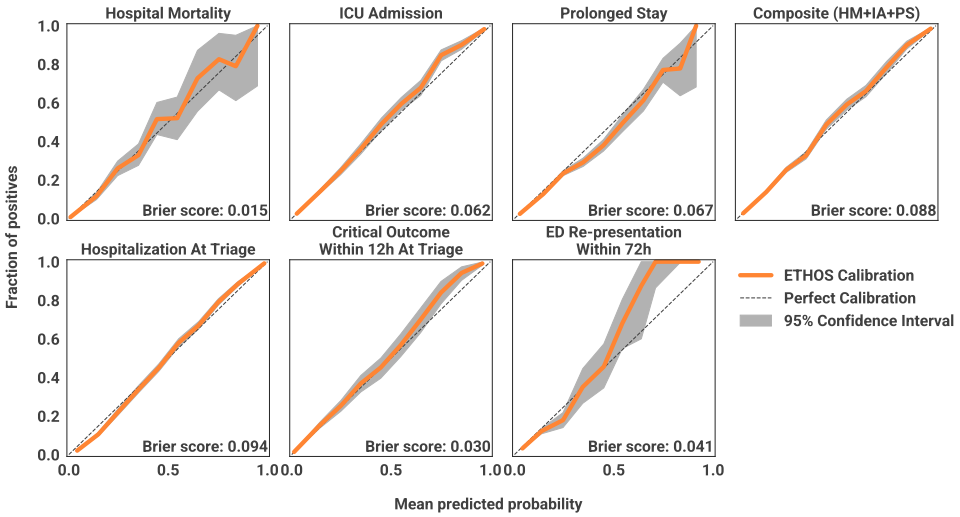


Figure 4.7: **Calibration across seven tasks.** Predicted probabilities versus observed outcome rates, binned by deciles. Shaded regions denote 95% confidence intervals. ARES exhibits excellent or good calibration for all endpoints, with minor overestimation only at the highest risk decile for hospitalization at triage.

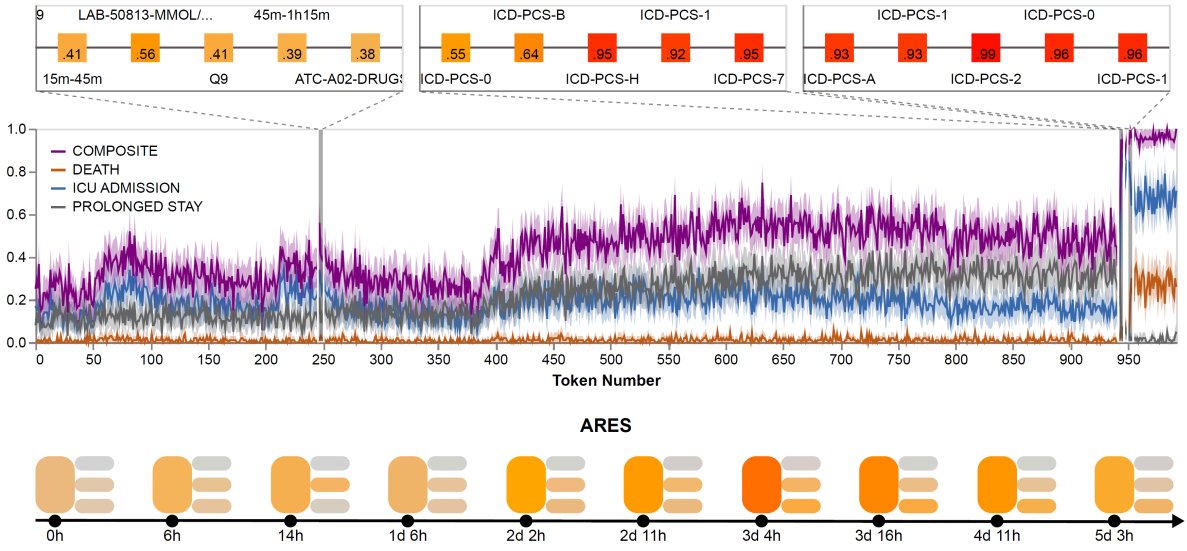


Figure 4.8: **Personalized risk trajectories and salient events.** Lower panel: evolving probabilities for death, ICU admission, prolonged stay, and composite risk, with 95% confidence bands derived from Monte Carlo sampling. Upper panel: influential token regions identified by attribution analysis, including high-quantile laboratory results and respiratory interventions.

5. Federated Timeline Synthesis

This chapter is based on the preprint [39], entitled *Federated Timeline Synthesis: Scalable and Private Methodology For Model Training and Deployment*. The Federated Timeline Synthesis (FTS) framework extends the ETHOS architecture to address one of the most pressing challenges in healthcare AI: reconciling the need for large-scale, heterogeneous data with the equally important requirement of strict privacy protection and institutional autonomy. Modern clinical AI systems thrive on large datasets; however, the sensitive nature of patient data, coupled with regulatory and infrastructural barriers, makes the centralized pooling of electronic health records (EHRs) largely infeasible. FTS proposes a principled solution by introducing a methodology for generating synthetic, privacy-preserving Patient Health Timelines (PHTs) that retain both the statistical fidelity and temporal coherence of real-world data. In doing so, FTS makes it possible to train and deploy foundation models across multiple institutions without direct data sharing.

5.1. Motivations and Related Works

5.1.1. Challenges in Scaling Foundation Models for Healthcare

The remarkable success of foundation models in natural language processing, exemplified by the GPT family of models, has set expectations for similarly transformative breakthroughs in healthcare. However, direct translation of these methods into the clinical domain is not straightforward. Clinical data are fragmented across institutions, heavily regulated under frameworks such as GDPR and CCPA, and inherently heterogeneous. Patient populations vary in demographics, disease prevalence, and medical practices; documentation and coding differ substantially across healthcare systems. Moreover, electronic health records are noisy, sparse, and multimodal, comprising structured codes, laboratory values, vital signs, imaging, and unstructured clinical narratives.

These factors impose severe constraints on attempts to centralize data for training large-scale foundation models. Even when data aggregation is technically possible, issues of fairness, generalizability, and representativeness remain unresolved. FTS was designed to tackle these barriers by enabling collaboration without the exchange of raw data. Its central innovation lies in leveraging tokenized PHTs as a universal abstraction of longitudinal patient histories. Tokenization not only standardizes heterogeneous modalities into a discrete, language-like format, but also serves as a first level of anonymization. Locally generated synthetic timelines provide an additional privacy-preserving layer, ensuring that no identifiable information leaves the source institution (Figure 5.1).

5.1.2. Federated Learning

Federated learning (FL) emerged as a paradigm for decentralized model training, wherein model parameters are trained locally and aggregated centrally, thereby avoiding the direct sharing of raw data. The canonical algorithm, Federated Averaging (FedAvg) [31], demonstrated that iterative averaging of locally trained updates could converge despite heterogeneity in data distributions. Since then, the FL literature has expanded to address communication

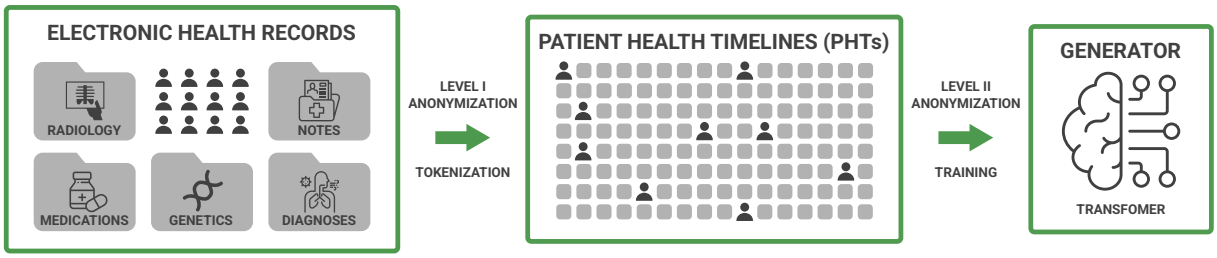


Figure 5.1: Two layers of anonymization in FTS: raw EHRs are first transformed into tokenized Patient Health Timelines (PHTs), which already abstract away direct identifiers. Local generative models then produce synthetic timelines, introducing a second protective layer before any information is shared across institutions.

efficiency, statistical non-IID data, robustness to adversarial clients, personalization, and privacy guarantees [24, 20, 61, 29, 17, 11].

Despite its promise, conventional FL presents challenges in healthcare. Communication costs remain high due to repeated gradient exchanges, training can diverge under strong data heterogeneity, and even weight sharing can expose subtle privacy risks. In addition, institutional variation in coding, outcome definitions, and data collection makes it difficult to define a uniform task space for federated model training.

5.1.3. Synthetic EHRs and Federated Synthesis

In parallel, synthetic EHR generation has been studied as a privacy-preserving alternative to data sharing. Early approaches, such as medGAN [10], demonstrated the feasibility of producing realistic, multi-label records. EHR-M-GAN [2] extended this to temporal and multimodal ICU data, while frameworks like EHR-Safe [60] integrated privacy-preserving mechanisms. More recent studies [49, 46, 64] confirmed that synthetic datasets can support downstream predictive modeling, mitigate fairness issues, and enable data sharing across institutions without exposing sensitive records.

Federated Synthesis (FS) extends this paradigm by combining generative modeling with distributed learning. In FS, clients generate synthetic data locally and share only these anonymized samples or trained generators with a central coordinator [53, 4, 27, 28]. This paradigm reduces communication costs, enhances privacy, and avoids the alignment overhead required by conventional FL.

5.1.4. Federated Timeline Synthesis

FTS builds upon these strands by combining federated learning, synthetic data generation, and tokenized PHTs into a single framework. Each institution trains an autoregressive transformer locally on its PHTs. Instead of exchanging gradients or raw data, institutions transmit model weights to a central server. The server then synthesizes a large corpus of timelines and uses it to train a **Global Generator (GG)**, which is redistributed back to client sites for local deployment (Figure 5.2).

This design achieves three advantages. First, it strengthens privacy guaranties: raw events and even synthetic trajectories remain local, while only generator weights are transmitted. Second, it ensures interoperability by using tokenized PHTs as a standardized abstraction across heterogeneous institutions. Third, it provides scalability: the Global Generator can be applied to diverse downstream tasks without retraining and can flexibly integrate new data modalities by expanding the token vocabulary.

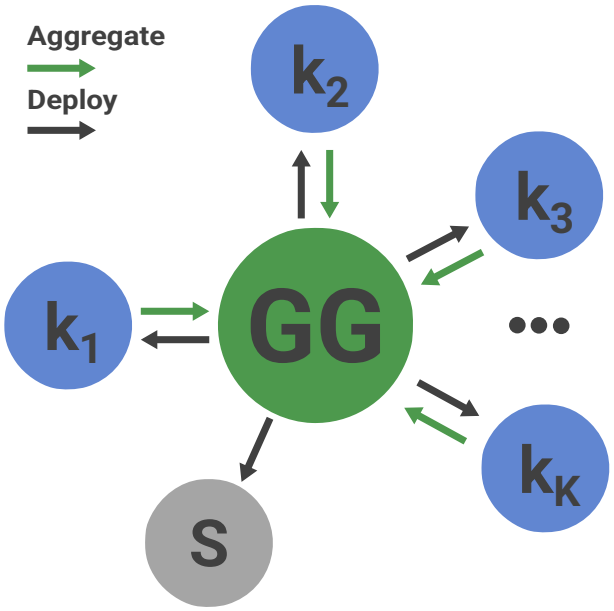


Figure 5.2: Federated Timeline Synthesis workflow. Clients train local generators on tokenized PHTs and send only model weights to a central server. The server constructs a Global Generator (GG) capable of synthesizing timelines at scale. The GG is redistributed to clients for downstream inference and synthetic cohort generation.

Significance of FTS

By converting EHRs into sequences of discrete tokens—including interval tokens for time gaps, quantile tokens for continuous values, and hierarchical tokens for structured codes—FTS achieves language-like modeling of healthcare timelines. This design ensures (i) privacy, as sensitive raw values never leave the source; (ii) interoperability, as site-specific structures are harmonized into a universal vocabulary; and (iii) extensibility, as additional modalities such as imaging, genomics, or wearable data can be encoded into the same timeline representation.

The Global Generator thus serves as a federated foundation model: it can support zero-shot inference, counterfactual simulation, and synthetic cohort creation across diverse institutions. Much like large language models in NLP, the GG enables downstream fine-tuning and modular adaptation, offering a blueprint for scalable, equitable, and privacy-preserving clinical AI.

5.1.5. Contributions

This work makes two primary contributions.

1. FTS introduces a novel integration of generative modeling with federated learning, leveraging PHTs to create a communication-efficient, privacy-preserving framework for foundation model training across institutions.

2. Through experiments on open-source structured EHR data, FTS demonstrates that models trained on synthetic PHTs achieve performance comparable to those trained on real centralized data. This establishes synthetic timelines as a viable surrogate for real-world records in downstream tasks, paving the way for scalable, privacy-conscious clinical AI.

5.2. Methods

Federated Timeline Synthesis operates on tokenized Patient Health Timelines to ensure safety, privacy, and computational efficiency. The mathematical formalism of the Patient Health Timeline representation and zero shot inference follows the formulation introduced in [40] and is presented in Section 5.2.1. The training and deployment workflow for Federated Timeline Synthesis, including the role of the Global Generator, was outlined in the preceding sections and is summarized conceptually in Figures 5.1 and 5.2. Throughout, a two layer anonymization pipeline, tokenization and generative modeling, underpins privacy preservation.

5.2.1. Medical Data Representation and Inference

Timeline representation. Each patient p is represented by a strictly ordered sequence of clinical events,

$$\mathcal{T}_p \;=\; \big(e_{p,1},\, e_{p,2},\, \ldots,\, e_{p,N_p}\big), \qquad e_{p,i} = \big(\tau_{p,i},\, y_{p,i}\big),$$

with the temporal keys satisfying lexicographic order

$$(\tau_{p,1},\, s_{p,1}) <_{\text{lex}} (\tau_{p,2},\, s_{p,2}) <_{\text{lex}} \cdots <_{\text{lex}} (\tau_{p,N_p},\, s_{p,N_p}),$$

where $\tau_{p,i} \in \mathbb{R}$ denotes a timestamp and $s_{p,i} \in \{1, 2, \dots\}$ breaks ties for coincident times using a deterministic secondary key, for example, an index in a sorted codebook. The payload

$$y_{p,i} \in \bigcup_{m \in \{\text{scalar, vector, text, image,}\dots\}} \mathcal{Y}_m \quad \text{or} \quad y_{p,i} = \emptyset$$

captures heterogeneous modalities such as laboratory measurements, vital signs, clinical notes, or images, with $\emptyset$ indicating events without an attached payload.

Patient Health Timeline. A tokenization map T converts each event $e_{p,i}$ into a short subsequence of tokens from a vocabulary $\mathcal{V}$:

$$T:\; e_{p,i} \mapsto (x_{p,j},\, x_{p,j+1},\, \ldots,\, x_{p,j+k_i-1}), \qquad x_{p,\cdot} \in \mathcal{V},$$

typically emitting between one and ten tokens per event. For example, an event occurring approximately six minutes after the previous one with ICD 10 CM code `E11.65` can be tokenized as

$$T(e_{p,i}) = \big(\underbrace{\texttt{INT}_{5\text{min}}}_{\Delta\tau},\; \underbrace{\texttt{E11}}_{\text{ICD prefix}},\; \underbrace{\texttt{65}}_{\text{ICD suffix}}\big).$$

Concatenating tokenizations yields the patient level sequence

$$\mathbf{x}_p = (x_{p,1},\, x_{p,2},\, \ldots,\, x_{p,L_p}),$$

and the corpus $\{\mathbf{x}_p\}_{p=1}^{P}$ is used for autoregressive training. This implements the first anonymization layer in Figure 5.1 by abstracting raw timestamps and magnitudes into discrete symbols.

Token classes. The vocabulary $\mathcal{V}$ is partitioned into four functional classes.

Static tokens. Patient level attributes are emitted once at the sequence start; for example, the age bin at the timeline start, sex, baseline comorbidities, marital status, or socioeconomic markers. These fixed position tokens stabilize conditioning for downstream generation.

Hierarchical tokens. High cardinality codes are decomposed into ordered fragments that reflect taxonomy levels; for example, `E11.65` $\rightarrow$ `E11` $\rightarrow$ `65` is used for ICD 10 CM, and similarly for ATC medications, CPT, and

ICD 10 PCS procedures. This preserves specificity while exposing the compositional structure that is useful for generalization and interoperability.

Interval tokens. Inter event gaps $\Delta\tau_i = \tau_{p,i} - \tau_{p,i-1}$ are discretized into B nominal durations, for example, 5 minutes, 20 minutes, 1 hour, and 1 day, producing $\texttt{INT}_b$. If $\Delta\tau_i$ lies below a minimum threshold, typically half of the shortest bin, no interval token is emitted.

Measurement, quantile tokens. Continuous values are mapped to deciles using the empirical cumulative distribution function F:

$$q = \min\big(\lfloor F(v)\, Q \rfloor,\, Q-1\big), \quad Q = 10, \qquad \text{emit } \texttt{QNT}_q.$$

For a blood pressure event recorded below the minimal interval bin, the tokenization may be

$$T(e_{p,i}) = \big(\texttt{BP},\ \texttt{QNT}_5 \text{ (systolic)},\ \texttt{QNT}_7 \text{ (diastolic)}\big).$$

Multimodal embeddings. Unstructured or high dimensional payloads $y_{p,i}^{(m)}$ are embedded via pretrained encoders,

$$\mathbf{z}_{p,i}^{(m)} = h_m\big(y_{p,i}^{(m)}\big) \in \mathbb{R}^d, \quad m \in \{\text{notes}, \text{images}, \text{genomics}\},$$

and interleaved at the appropriate positions. Discrete tokens $x_{p,j}$ are mapped via a learned embedding $E : \mathcal{V} \to \mathbb{R}^d$, $\mathbf{e}_{p,j} = E(x_{p,j})$. The final input is the chronologically ordered sequence of $\{\mathbf{e}_{p,j}\}$ and $\{\mathbf{z}_{p,i}^{(m)}\}$.

Zero shot probabilistic inference via simulated futures. Given an observed prefix $\mathbf{x}_{p,1:L_p}$, a trained generator f_{θ^*} samples N future timelines

$$\{\tilde{\mathbf{x}}_p^{(n)}\}_{n=1}^N \sim f_{\theta^*}\big(\cdot \,\big|\, \mathbf{x}_{p,1:L_p}\big),$$

stopping at task specific criteria, for example, discharge, ICU admission, death, or a fixed horizon. For a binary event $\mathcal{E}$,

$$\widehat{P}(\mathcal{E} \mid \mathbf{x}_{p,1:L_p}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{\mathcal{E}\text{-token} \in \tilde{\mathbf{x}}_p^{(n)}\}.$$

For multiclass outcomes with classes $c \in \{1, \ldots, C\}$,

$$\widehat{P}(\mathcal{E} = c \mid \mathbf{x}_{p,1:L_p}) = \frac{M_c}{N}, \qquad \sum_c M_c = N.$$

For regression targets,

$$\widehat{v}_p = \frac{1}{N}\sum_{n=1}^{N} v_n,$$

where v_n is extracted from the n th simulated timeline. This Monte Carlo procedure natively handles competing risks and yields calibrated uncertainty by construction, aligning with the zero shot design in [40]. It constitutes the second anonymization layer in Figure 5.1 by generating synthetic continuations rather than revealing raw data.

5.2.2. Federated Timeline Synthesis Framework

Local training. Assume K institutions with disjoint PHT corpora $\texttt{PHT}_k$. Each client trains an autoregressive generator f_{θ_k} with next token prediction,

$$\mathcal{L}_k(\theta_k) = -\sum_{p \in \texttt{PHT}_k} \sum_{j=1}^{L_p} \log p_{\theta_k}(x_{p,j} \mid x_{p,1:j-1}),$$

using the standardized vocabulary and tokenization rules described above. Training is confined to the local environment; raw events and intermediate batches are never exported.

Synthesis and aggregation by distillation. Upon convergence, clients transmit parameters $\{\theta_k\}_{k=1}^{K}$ to a coordinator. Rather than averaging gradients or weights under a federated optimization routine, the coordinator draws large synthetic corpora from each local model,

$$\widetilde{\texttt{PHT}} = \bigcup_{k=1}^{K} \{\tilde{\mathbf{x}}_{k,i}\}_{i=1}^{M}, \qquad \tilde{\mathbf{x}}_{k,i} \sim f_{\theta_k}(\cdot),$$

optionally conditioning on static tokens to balance strata of interest, for example, age or sex. A **Global Generator** f_{θ^*} is then trained on $\widetilde{\texttt{PHT}}$ via the same objective,

$$\mathcal{L}_{\text{syn}}(\theta) = -\sum_{\tilde{\mathbf{x}} \in \widetilde{\texttt{PHT}}} \sum_{j=1}^{|\tilde{\mathbf{x}}|} \log p_\theta(\tilde{x}_j \mid \tilde{x}_{1:j-1}),$$

thereby distilling multi-site variability into a single, scalable generator without exposing raw records.

Redistribution and deployment. The resulting f_{θ^*} is redistributed to contributors and new sites. Local deployment supports zero shot inference, synthetic cohort creation, and token space extensions; for example, adding site specific codes or new modalities without retraining from scratch. Because only model parameters and synthetic sequences are exchanged, the workflow remains privacy preserving and communication efficient, consistent with Figures 5.1 and 5.2.

5.2.3. Downstream Tasks

Performance was evaluated across five clinically meaningful tasks using zero shot inference:

1. **Diagnosis Related Group prediction, multiclass.** A discharge time token is used as a stopping criterion, and the most probable DRG token is emitted. In the Medical Information Mart for Intensive Care dataset, this corresponds to an approximately eight hundred class classification problem.

2. **Sequential Organ Failure Assessment score regression.** The first day Sequential Organ Failure Assessment score is estimated from historical context by mapping predicted quantile tokens to expected scores.

3. **Thirty day readmission, binary.** Generation starts at discharge and proceeds forward in time. New admission or death within thirty days is treated as a positive outcome.

4. **ICU admission, binary.** Generation starts at hospital admission; admission to the intensive care unit or in hospital death counts as a positive outcome.

5. **In hospital mortality, binary.** Generation starts at hospital admission; only death during the stay counts as a positive outcome.

5.2.4. Overall Score Computation and Confidence Intervals

To provide a single, interpretable ranking across the five metrics, a global score S_i is computed for each method i as an inverse variance weighted sum of Min–Max normalized metric values. Each metric's variance is estimated from its reported ninety five percent confidence interval, so that metrics with tighter uncertainty contribute more to the overall score.

Let $m_{i,k}$ denote the observed value of the metric k for the method i, with a reported ninety five percent confidence interval $[\, m_{i,k}^{\text{low}},\, m_{i,k}^{\text{high}}]$. The global score S_i and its ninety five percent confidence interval are obtained as follows.

Standard error and variance for each metric:

$$h_{i,k} = \frac{m_{i,k}^{\text{high}} - m_{i,k}^{\text{low}}}{2}, \qquad \sigma_{i,k} = \frac{h_{i,k}}{1.96}, \qquad \text{Var}(m_{i,k}) = \sigma_{i,k}^2. \tag{5.1}$$

Define the per metric range,

$$m_{(1),k} = \min_i m_{i,k}, \qquad m_{(N),k} = \max_i m_{i,k}.$$

Normalization:

$$\hat{m}_{i,k} = \frac{m_{i,k} - m_{(1),k}}{m_{(N),k} - m_{(1),k}} \in [0,1], \tag{5.2}$$

with variance

$$\text{Var}(\hat{m}_{i,k}) = \frac{\sigma_{i,k}^2}{\left(m_{(N),k} - m_{(1),k}\right)^2}. \tag{5.3}$$

Inverse variance weights:

$$w_{i,k} = \frac{1/\text{Var}(\hat{m}_{i,k})}{\sum_{\ell=1}^{M} 1/\text{Var}(\hat{m}_{i,\ell})}, \qquad \sum_{k=1}^{M} w_{i,k} = 1. \tag{5.4}$$

Global score and uncertainty:

$$S_i = \sum_{k=1}^{M} w_{i,k}\, \hat{m}_{i,k}, \tag{5.5}$$

and, under independence across metrics,

$$\text{Var}(S_i) = \sum_{k=1}^{M} w_{i,k}^2 \,\text{Var}(\hat{m}_{i,k}) = \frac{1}{\sum_{k=1}^{M} 1/\text{Var}(\hat{m}_{i,k})}. \tag{5.6}$$

The ninety five percent confidence interval for S_i is then given by $S_i \pm 1.96\,\sqrt{\text{Var}(S_i)}$.

5.3. Experiments and Results

This section evaluates Federated Timeline Synthesis across five clinically meaningful downstream tasks, defined earlier in Section 5.2.3: Diagnosis Related Group assignment, Sequential Organ Failure Assessment score estimation, thirty day readmission, admission to the intensive care unit, and in hospital mortality. All experiments are conducted on MIMIC IV v2.2 [18, 19] with patient level splits to prevent information leakage across sets: `orig` (ninety percent), `test` (ten percent), `val1` (five percent), and `val2` (five percent). The evaluation proceeds in four stages designed to mirror realistic development and deployment: (1) division of the available training data in `orig` to emulate institutions of different sizes, (2) selection of the zero shot inference temperature that governs stochastic timeline simulation, (3) tuning of the generation temperature used to synthesize Patient Health Timelines, and (4) assessment of deployment scenarios that rely on a Global Generator for cross site collaboration. Stages (1) and (2) use only `val1` and `val2` to avoid test set peeking, whereas stages (3) and (4) are evaluated on `test`.

Model and training configuration. All generators follow a decoder only transformer design in the style of generative pretraining. To ensure transparency and reproducibility, capacity is fixed across comparisons: three transformer blocks, a hidden dimension of seven hundred sixty eight, twelve attention heads, a dropout of zero point three, and a two thousand forty eight token context window. Optimization uses AdamW with a learning rate that is cosine decayed from 6×10^{-4} to 1×10^{-5} over fifty thousand iterations. Each model trains for three hundred epochs, and the checkpoint with the lowest average of the last five validation losses is retained to reduce variance from stochastic optimization. The effective batch size is five hundred twelve. Training is executed on nodes equipped with eight NVIDIA A100 SXM four forty gigabyte graphics processing units and one terabyte of random access memory; wall clock times range from approximately four to thirty hours as a function of dataset size.

Stage 1: Division of training data and the effect of scale. To emulate institutions with unequal data resources, `orig` is progressively subsampled, and the multi task performance is evaluated on `val1` using the global score defined in Section 5.2.4. The results, visualized in Figure 5.3, reveal clear elasticity with respect to data scale. Performance remains stable down to approximately eighty percent of `orig`, degrades at twenty percent, and declines further at ten percent. These breakpoints motivate three regimes used consistently in subsequent experiments: `big` (eighty percent of `orig`), `small` (twenty percent), and `little` (ten percent, a strict subset of `small`). The monotone trend is not confined to a single endpoint; rather, it is visible across Diagnosis Related Group classification, Sequential Organ Failure Assessment regression, and the three binary outcomes. This stage establishes realistic baselines for low resource and mid resource sites.

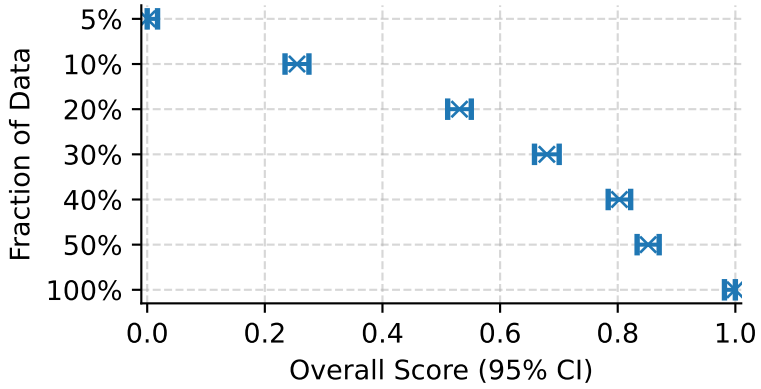


Figure 5.3: Stage 1. Overall downstream score as a function of the fraction of training data drawn from `orig`. Performance is resilient to moderate reductions, with marked decline at twenty percent and ten percent of the original pool. These inflection points motivate the `big`, `small`, and `little` regimes used in later stages.

Stage 2: Selection of the inference temperature for zero shot simulation. Zero shot inference proceeds by Monte Carlo simulation of future Patient Health Timelines. The sampling temperature balances exploration against exploitation when generating token sequences. Temperatures from zero point seven to one point two are swept using a model trained on `orig` and evaluated on `val2`. Figure 5.4 summarizes discrimination across tasks, and Figure 5.5 compares probability calibration for the three best settings on the binary endpoints. All three candidate settings yield well calibrated probabilities, with zero point nine delivering the best global score. The value of zero point nine is therefore adopted for all subsequent evaluations to standardize inference conditions.

Stage 3: Tuning generation temperature for synthetic timeline creation. Knowledge transfer from models trained on original records is effected by autoregressively generating synthetic Patient Health Timelines in the same token space. The generation temperature is hypothesized to govern a utility versus diversity trade off: lower temperatures concentrate probability mass on high confidence patterns and may improve short horizon fidelity, whereas higher temperatures yield more diverse sequences that can enhance robustness and coverage at the risk of occasional context violations. For each data regime, `big`, `small`, and `little`, synthetic counterparts are produced at temperatures $\{0.7, 0.9, 1.0, 1.1\}$, matching patient counts and demographic strata to their real sources. New models are then trained from scratch on each synthetic set and evaluated on `test` using a fixed inference temperature of zero point nine. Figure 5.6 provides an aggregate view, and Figure 5.7 shows calibration for the top settings.

Three observations emerge. First, a temperature of one point zero consistently yields the best global score across regimes, suggesting that the baseline token distribution of trained local generators already strikes an effective balance

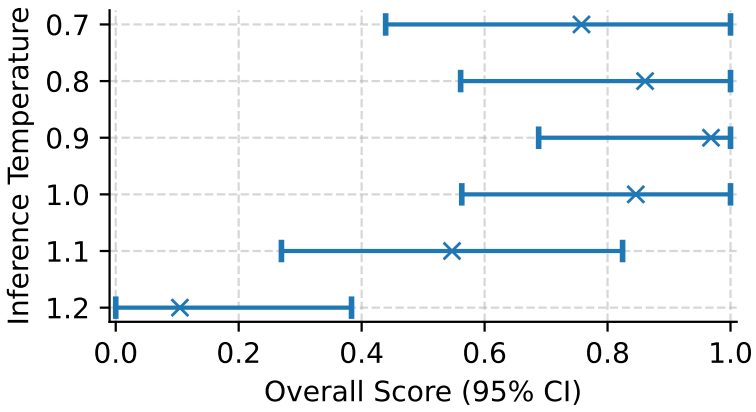


Figure 5.4: Stage 2. Discrimination as a function of inference temperature for zero shot timeline simulation. The temperature of zero point nine achieves the most favorable balance across the five downstream tasks and is used thereafter.

for downstream utility. Second, neighboring temperatures of zero point nine and one point one are competitive and exhibit similar calibration, indicating the robustness of the synthesis pipeline. Third, moving away from one point zero shifts token group frequencies and increases rare out of context emissions, especially for hierarchical codes and long interval tokens. Token frequency analyzes are summarized in Table 5.1. In practice, a simple default is supported: generate at temperature one point zero and adjust only when explicit coverage or privacy constraints motivate additional diversity.

Stage 4: Deployment scenarios for Federated Timeline Synthesis. Two deployment patterns are examined to reflect realistic cross institutional collaboration. In the first scenario, multiple institutions train local generators on their Patient Health Timelines and contribute only model parameters. A central coordinator aggregates these to train a Global Generator which, in turn, synthesizes timelines at scale for downstream modeling. In the second scenario, a single institution adopts the Global Generator for direct inference and optionally augments local training with Global Generator synthesized data. A spectrum of data budgets and mixtures of real and synthetic datasets is simulated to quantify how much synthetic augmentation can substitute for scarce local data.

The summary in Figure 5.8 shows a coherent picture across all five tasks. Synthetic augmentation substantially benefits low resource regimes. Training on `small` plus Global Generator synthetic data approaches the `big` baseline, and `little` plus Global Generator synthetic data exceeds the performance of `small` alone by a comfortable margin. Performance plateaus once training data exceed the `small` threshold, consistent with the small gap between `big` and `big` augmented with `small` synthetic data. Training solely on synthetic data underperforms training on original records; for example, `big_synth` versus `big`, indicating that synthetic timelines preserve most, but not all, higher order dependencies. In practical terms, Federated Timeline Synthesis is most impactful where data limitations and privacy constraints preclude pooling. In these settings, Global Generator based augmentation narrows the gap to well resourced sites without exposing patient level information.

5.4. Computational Cost of Federated Timeline Synthesis

Federated Timeline Synthesis (FTS) offers notable computational efficiency relative to conventional federated learning and privacy preserving training frameworks. Traditional federated approaches typically require iterative gradient exchanges and frequent synchronization across clients, incurring substantial communication and

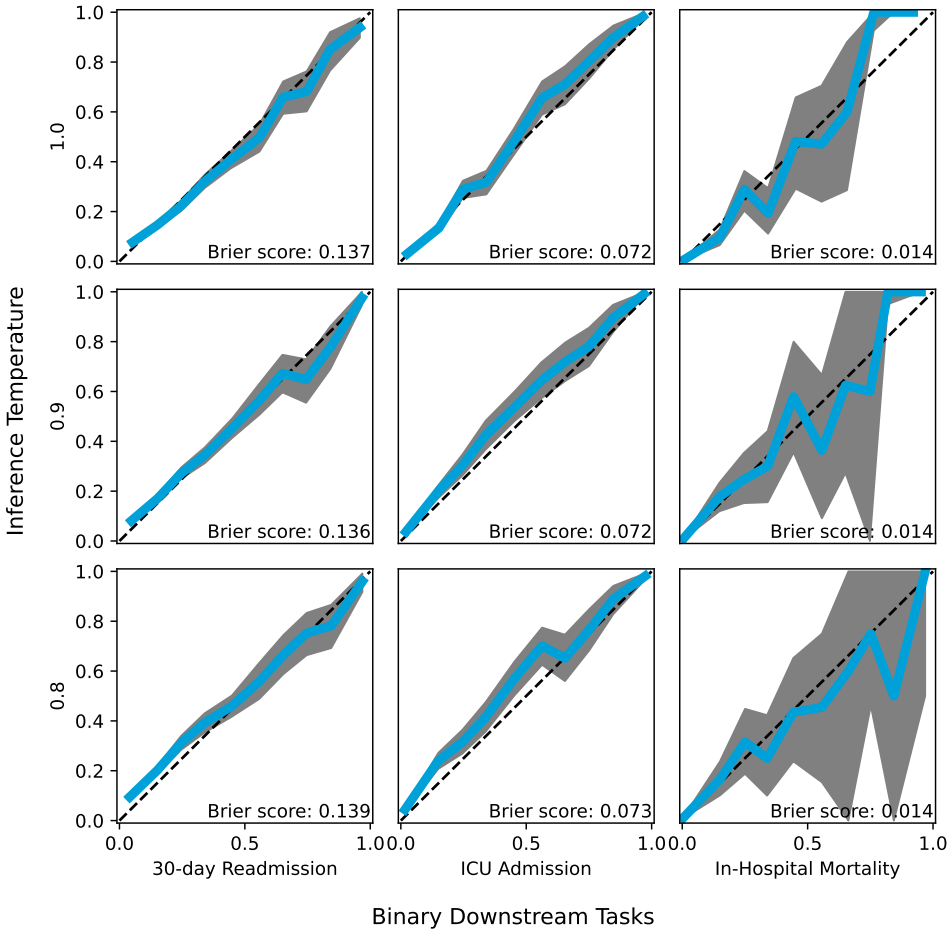


Figure 5.5: Calibration for the top inference temperatures on the binary tasks. At temperature zero point nine, predicted probabilities closely track the identity line while maintaining discrimination, indicating a favorable bias variance trade off for zero shot risk estimation.

orchestration overhead. In contrast, FTS transmits trained generator weights once per participating site during coordination, thereby reducing network traffic and simplifying scheduling. This one shot coordination is especially advantageous when client connectivity is intermittent or when regulatory processes constrain the cadence of cross site communication.

From an execution perspective, model training is dominated by local autoregressive pretraining on tokenized Patient Health Timelines (PHTs), which scales linearly with token count and batch size under the fixed capacity configuration described earlier. Central aggregation and Global Generator construction are comparatively lightweight, as they occur less frequently than local update steps and avoid repeated gradient synchronization.

The principal additional cost introduced by FTS appears at inference time. Accurate zero shot prediction is obtained by sampling multiple future PHTs per patient to estimate outcome probabilities. This Monte Carlo procedure produces calibrated and scenario aware estimates, but it increases per patient latency in proportion to the number of samples. Two observations mitigate this overhead. First, inference is performed once per evolving timeline state and can be amortized across multiple downstream tasks, since the same set of simulated futures supports mortality, ICU admission, readmission, and other endpoints without retraining. Second, sampling can be adaptively budgeted: fewer draws are sufficient when predicted risk is far from decision thresholds, whereas additional draws are reserved for borderline cases to stabilize decisions. In practice, the combination of amortization across tasks and adaptive sampling yields acceptable throughput on modern accelerators and aligns with the ongoing

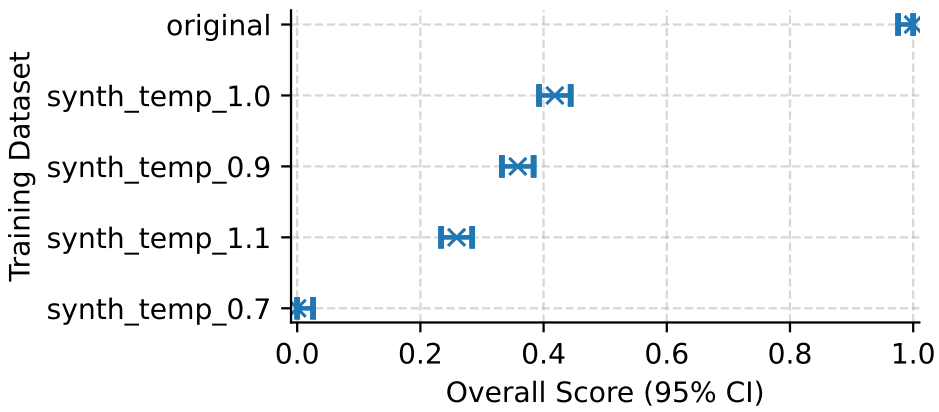


Figure 5.6: Stage 3. Effect of synthesis temperature on downstream performance for synthetic datasets that mirror the size and demographic structure of their real counterparts. The best overall performance occurs at temperature one point zero, with stable calibration for zero point nine through one point one.

trend toward a lower unit cost of compute.

5.5. Fidelity Evaluation

Synthetic timelines should preserve the statistical regularities that drive downstream utility while avoiding memorization of identifiable records. To quantify the alignment between real and synthetic data, two complementary fidelity metrics are adopted that do not depend on visit based tokenization and therefore match the timeline setting.

Unigram code distribution (R^2). Let f_i^{real} denote the marginal frequency of code i in the real dataset and f_i^{synth} the corresponding frequency in the synthetic dataset. The coefficient of determination

$$R^2_{\text{Unigram}} = 1 - \frac{\sum_i \left(f_i^{\text{real}} - f_i^{\text{synth}}\right)^2}{\sum_i \left(f_i^{\text{real}} - \bar{f}^{\text{real}}\right)^2} \tag{5.7}$$

measures preservation of marginal code usage. High R^2_{Unigram} indicates that synthetic corpora reproduce population level code mix, an essential property for training discriminative models on synthetic data.

Dimension wise correlation (R^2). To assess patient level composition, for each patient p define the normalized code frequency vector

$$\mathbf{v}_p = \frac{\text{code counts for patient } p}{\text{total codes for patient } p}. \tag{5.8}$$

Averaging across patients produces $\bar{\mathbf{v}}^{\text{real}}$ and $\bar{\mathbf{v}}^{\text{synth}}$. The coefficient of determination

$$R^2_{\text{DimWise}} = 1 - \frac{\sum_i \left(\bar{v}_i^{\text{real}} - \bar{v}_i^{\text{synth}}\right)^2}{\sum_i \left(\bar{v}_i^{\text{real}} - \bar{v}^{\text{real}}\right)^2} \tag{5.9}$$

captures agreement in the average per patient code composition. High R^2_{DimWise} suggests that synthetic timelines reflect the typical distribution of codes within a patient, not just aggregate counts.

Results and analysis

Tables 5.3 and 5.4 summarize fidelity on two evaluation sets, referred to as the timeline and readmission datasets, across sampling temperatures and data scales. Several consistent patterns emerge and are visually supported by the `big` dataset unigram comparisons in Figure 5.9.

First, under `big` and `small` regimes, a temperature of one point zero achieves near perfect alignment for both Unigram and DimWise metrics on the timeline dataset ($R^2 \geq 0.998$ in Table 5.3) and on the readmission dataset ($R^2 \geq 0.996$ for `big`, $R^2 \approx 0.996$ for `small` in Table 5.4). This indicates that when the local generator is trained on sufficient data, synthesis at the default temperature preserves both population level code mix and average per patient composition. Second, temperatures zero point nine and one point one remain competitive in these regimes, with differences residing at the third decimal place; the real versus synthetic probability scatter in Figure 5.9 closely hugs the identity line for zero point nine through one point one, confirming robustness in the stable temperature band.

Third, when data are scarce (`little`), fidelity decreases, especially for readmission, where Unigram R^2 falls into the 0.80–0.86 range depending on temperature. DimWise R^2 remains comparatively higher in the same setting, suggesting that the average within patient composition is better preserved than marginal frequencies when the training pool is small. This divergence aligns with downstream results in Table 5.2, where models trained only on `little_synth` underperform larger synthetic or real sets. Fourth, modestly increasing temperature can aid rare code coverage in the `little` setting: for readmission, one point one improves Unigram and DimWise relative to zero point nine (Table 5.4, 0.954 and 0.979 versus 0.808 and 0.934), indicating that a slight entropy increase counteracts undercoverage from limited data.

Finally, fidelity maxima coincide with downstream utility in Stage 3 (Figure 5.6), where temperature one point zero yields the strongest overall score and neighboring settings are close. The same settings also preserve calibration (Figure 5.7). These concordant trends support a simple operational choice: synthesize at one point zero by default and consider a small temperature increase only in extremely data limited contexts to improve rare code representation.

Implications. The fidelity results complement the downstream findings in Stage 3 and Stage 4. Temperatures that maximize R^2 also maximize discrimination and preserve calibration, supporting a single default for synthesis. In extremely data limited settings, a minor temperature increase can recover diversity for rare codes without materially degrading calibration, consistent with the narrow spread observed in Figure 5.7. The consistently high DimWise scores indicate that synthetic timelines maintain a realistic per patient composition, explaining why models trained solely on synthetic data retain a substantial fraction of utility, even when marginal distributions are imperfectly matched.

5.6. Privacy Preservation and Fidelity of Synthetic Data

Predictive accuracy alone is insufficient for clinical deployment. FTS is designed around two complementary privacy abstractions. First, raw EHRs are transformed locally into tokenized PHTs that replace exact timestamps and magnitudes with discrete, standardized symbols, including interval tokens for time gaps and quantile tokens for continuous values. This mapping removes direct identifiers and reduces the risk of linkage via raw timestamps or rare numeric patterns. Second, collaboration proceeds through stochastic sequence generation and, where coordination is required, through the exchange of model parameters rather than records. Sites, therefore, share neither raw events nor exact patient trajectories during training or evaluation.

A practical advantage of synthetic generation is the tunability of the realism privacy trade off. Institutions can tighten privacy by increasing synthesis temperature within the empirically stable range, imposing sampling constraints on rare hierarchical codes, or conditioning on broad demographic strata to control cohort composition without emitting unique record fragments. Conversely, when local policy emphasizes fidelity for a particular project, the temperature can be kept near one point zero and constraints can be relaxed while still avoiding raw data export. This flexibility allows FTS to adapt to heterogeneous regulatory contexts while maintaining utility for downstream modeling.

5.6.1. Summary

The results across stages and tasks establish the viability and value of Federated Timeline Synthesis.

- **Scale sensitivity (Stage 1).** Performance was resilient to moderate reductions in training data and declined at twenty percent and ten percent of `orig` (Figure 5.3), motivating the `big`, `small`, and `little` regimes used throughout.
- **Zero shot inference configuration (Stage 2).** A fixed inference temperature of zero point nine balanced discrimination and calibration across all five downstream tasks (Figures 5.4 and 5.5), and was adopted for all subsequent evaluations.
- **Synthesis configuration (Stage 3).** Generation at a temperature of one point zero maximized downstream utility with stable calibration (Figure 5.6); neighboring values of zero point nine and one point one were competitive and preserved fidelity (Figure 5.7).
- **Deployment patterns and augmentation (Stage 4).** Global Generator augmentation substantially improved low resource regimes, with `small` plus synthetic approaching `big`, and `little` plus synthetic exceeding `small` (Figure 5.8). Training solely on synthetic data underperformed compared to training on original records but retained a large fraction of utility across DRG classification, SOFA regression, readmission, ICU admission, and mortality (Table 5.2).
- **Computational profile.** Communication costs and orchestration were reduced compared to iterative federated learning, as only the trained generator weights were transmitted once per site. Inference incurred additional sampling costs, which were amortized across tasks and were controllable via adaptive sampling (Section 5.4).
- **Fidelity and privacy.** Unigram and dimension wise R^2 metrics indicated strong preservation of aggregate and per patient code statistics within the stable synthesis range (Section 5.5). Privacy was strengthened by tokenization and by training through sequence generation and parameter exchange rather than record sharing.

Taken together, these findings show that FTS enables the training and deployment of timeline based foundation models whose downstream behavior approaches that of models trained on centralized real data while providing materially stronger privacy guaranties. The approach preserves clinically salient signals and probabilistic calibration across mortality, ICU admission, readmission, SOFA estimation, and DRG classification, narrows performance gaps for data limited institutions via Global Generator augmentation, and reduces coordination costs relative to conventional federated learning.

5.7. Discussion: Deployment and Interoperability

Federated Timeline Synthesis provides a practical path for the global deployment of timeline native foundation models across diverse healthcare systems. Because training proceeds without raw data sharing, institutions can

contribute to and benefit from model improvements regardless of local data sovereignty rules. Because all information is represented in a common token vocabulary, results are reproducible and naturally comparable across sites with differing coding practices and documentation styles. The Global Generator serves as a portable foundation that can be redistributed for local inference, synthetic cohort construction, or targeted fine tuning, accommodating new modalities by extending the vocabulary rather than rebuilding pipelines.

In this way, FTS addresses three longstanding barriers simultaneously: privacy, interoperability, and scale. It supports learning from heterogeneous international practices while remaining deployable within the operational and legal boundaries of individual institutions, bringing the field closer to trustworthy and widely accessible clinical artificial intelligence.

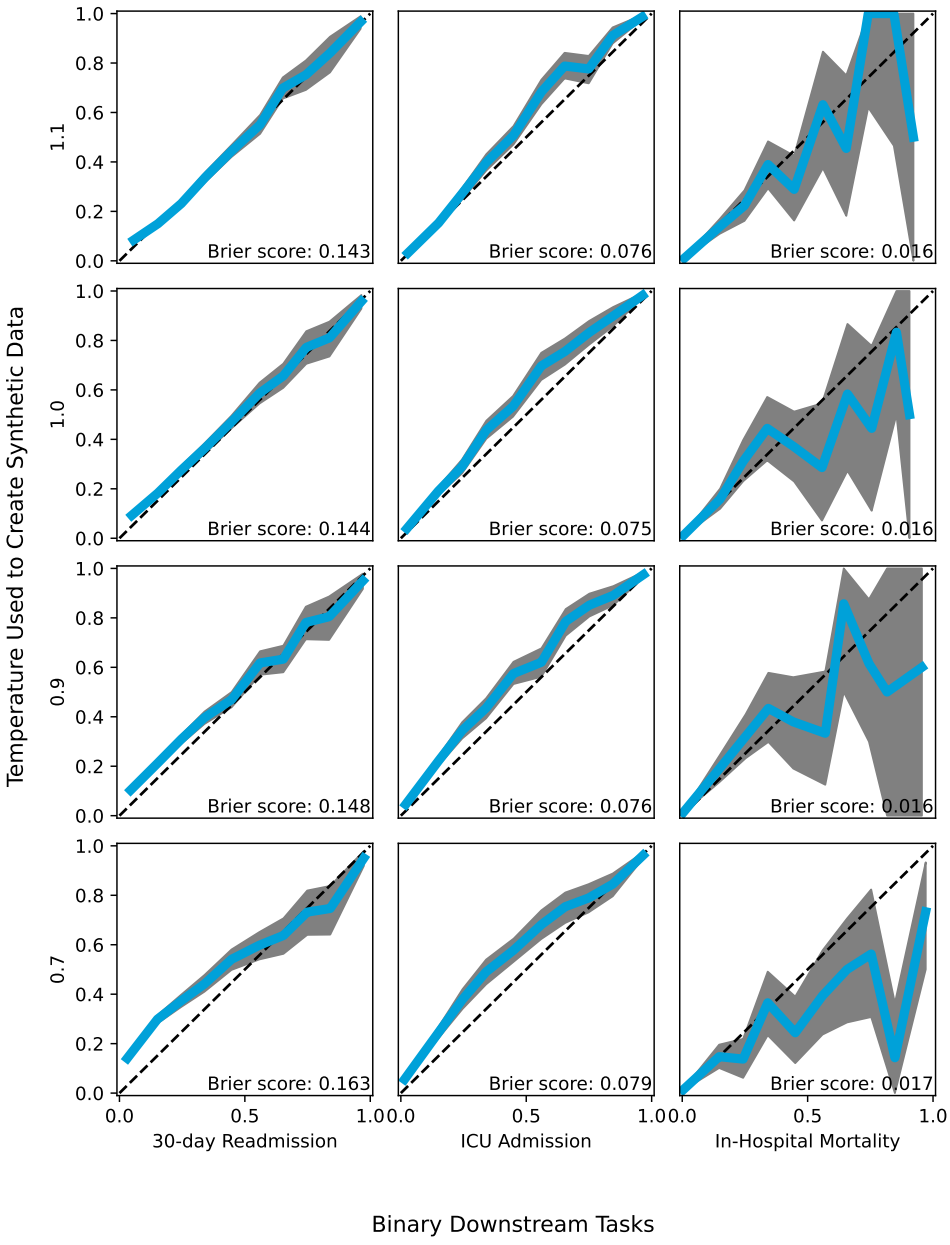


Figure 5.7: Calibration of models trained on synthetic datasets generated at top performing temperatures. Predicted probabilities closely follow the identity line, indicating that training on synthetic timelines preserves probabilistic calibration.

Code Group	Original		Synth_temp1		Synth_temp0.9		Synth_temp0.7		Synth_temp1.1	
	Count	N	Count	N	Count	N	Count	N	Count	N
LAB	72,174,268	200	71,106,715	200	59,794,628	200	60,518,244	200	93,198,749	200
ATC	20,858,757	87	22,795,722	83	12,452,636	83	6,101,998	77	37,949,275	86
ATC_4	20,858,744	12	22,795,876	12	12,452,591	11	6,101,190	11	37,950,129	12
ATC_SFX	20,769,779	208	22,705,112	195	12,394,075	184	6,075,559	157	37,832,338	205
Q1	9,494,082	1	8,861,399	1	7,572,188	1	10,245,103	1	11,804,981	1
Q2	8,621,537	1	8,588,113	1	7,134,201	1	6,742,542	1	11,356,270	1
Q3	8,288,665	1	8,283,252	1	6,984,193	1	6,550,402	1	10,732,378	1
Q4	7,616,529	1	7,553,914	1	6,398,353	1	5,986,846	1	9,714,675	1
Q5	7,601,185	1	7,594,698	1	6,539,317	1	6,199,183	1	9,592,139	1
Q7	7,285,786	1	7,300,422	1	6,344,285	1	6,242,105	1	9,191,142	1
Q6	7,213,924	1	7,237,419	1	6,231,882	1	5,928,606	1	9,150,938	1
Q8	6,755,991	1	6,784,355	1	5,858,041	1	5,732,405	1	8,685,969	1
Q9	6,510,118	1	6,482,137	1	5,702,253	1	6,048,333	1	8,233,991	1
ICD_CM	6,230,466	2,880	6,622,075	2,577	4,477,122	2,431	1,849,592	2,202	8,643,801	2,750
Q10	5,960,105	1	5,653,703	1	5,257,956	1	7,745,200	1	7,093,716	1
ICD_PCS	3,197,383	34	2,942,837	34	2,083,372	34	1,078,723	34	4,133,904	34
VITAL	1,560,547	1	1,589,742	1	2,097,787	1	3,442,893	1	1,129,540	1
1h15m-2h	1,532,311	1	1,595,616	1	953,791	1	438,114	1	2,514,137	1
3h-5h	1,481,319	1	1,401,654	1	954,219	1	649,646	1	1,930,502	1
2h-3h	1,468,528	1	1,467,322	1	912,352	1	455,639	1	2,186,806	1
15m-45m	1,400,879	1	1,524,672	1	850,931	1	383,343	1	2,749,798	1
BMI	1,190,022	10	1,134,606	11	1,583,297	11	2,725,486	11	794,648	11
45m-1h15m	1,147,674	1	1,218,088	1	686,810	1	280,946	1	2,086,068	1
5h-8h	911,451	1	841,629	1	594,188	1	385,975	1	1,110,288	1
5m-15m	907,753	1	1,026,894	1	519,283	1	210,106	1	1,989,240	1
8h-12h	797,169	1	741,521	1	789,513	1	1,206,250	1	769,878	1
TRANSFER	599,818	38	579,025	38	409,007	38	203,179	38	818,548	38
12h-18h	571,804	1	569,864	1	612,846	1	790,878	1	549,053	1
2mt-6mt	367,454	1	388,899	1	469,075	1	740,404	1	324,090	1
=6mt	350,714	1	320,753	1	375,142	1	489,934	1	271,160	1
30d-2mt	340,770	1	363,176	1	426,013	1	530,503	1	303,286	1
INSURANCE	310,529	3	291,693	3	233,662	3	128,742	3	332,829	3
HOSPITAL_DISCHARGE	310,529	1	295,194	1	237,341	1	130,433	1	325,726	1
DISCHARGE_LOCATION	310,529	10	295,409	10	237,408	10	130,470	10	326,145	10
HOSPITAL_ADMISSION	310,529	1	291,589	1	233,632	1	128,740	1	332,467	1
DRG	310,529	770	293,586	749	236,394	741	130,432	698	333,297	763
ADMISSION_TYPE	310,529	3	291,654	3	233,661	3	128,746	3	332,627	3
12d-20d	309,052	1	310,314	1	359,927	1	378,336	1	261.646	1
20d-30d	270,656	1	277,064	1	341,095	1	569,284	1	230,988	1
4d-7d	264,533	1	255,286	1	292,271	1	492,847	1	228,224	1
7d-12d	260,717	1	262,237	1	278,118	1	286,115	1	232,108	1
1d-2d	242,652	1	224,712	1	201,722	1	327,574	1	245,670	1
ED_REGISTRATION	212,943	1	199,513	1	159,303	1	87,303	1	226,059	1
ED_OUT	212,943	1	201,389	1	160,945	1	88,234	1	227,490	1
TIMELINE_END	192,773	1	192,773	1	192,773	1	192,773	1	192,773	1
TIMELINE_START	192,773	1	192,985	1	192,967	1	192,987	1	193,043	1
2d-4d	179,782	1	166,411	1	153,825	1	126,724	1	169,488	1
18h-1d	179,474	1	172,924	1	130,567	1	69,677	1	214,290	1
HCPCS	101,768	63	95,482	40	68,595	39	28,201	27	120,700	55
ICU_DISCHARGE	52,560	1	53,052	1	32,413	1	15,864	1	96,948	1
SOFA	52,560	1	51,917	1	31,963	1	16,068	1	95,691	1
ICU_ADMISSION	52,560	1	51,877	1	31,960	1	16,049	1	95,532	1
ICU_TYPE	52,560	9	51,894	9	31,962	9	16,059	9	95,654	9
MEDS_DEATH	21,022	1	22,423	1	15,050	1	7,515	1	34,980	1
GENDER	0	0	21	2	12	2	3	1	46	2
MARITAL	0	0	16	5	6	3	3	2	77	5
RACE	0	0	24	6	4	3	6	2	116	6
Total	238,780,034	4,367	242,612,649	4,017	183,998,923	3,845	165,768,512	3,525	339,736,051	4,232

Table 5.1: Token counts and number of unique tokens in each code subgroup for the original dataset and for synthetic datasets generated at temperatures 1.0, 0.9, 0.7 and 1.1. For each setting, the total token count (“Count”) and the corresponding unique-token count (“N”) are shown side by side. Note the unexpected presence of demographic tokens such as GENDER, MARITAL and RACE in the event timelines, and the higher frequency of TIMELINE_START compared to TIMELINE_END, both of which point to glitches in the synthetic data.

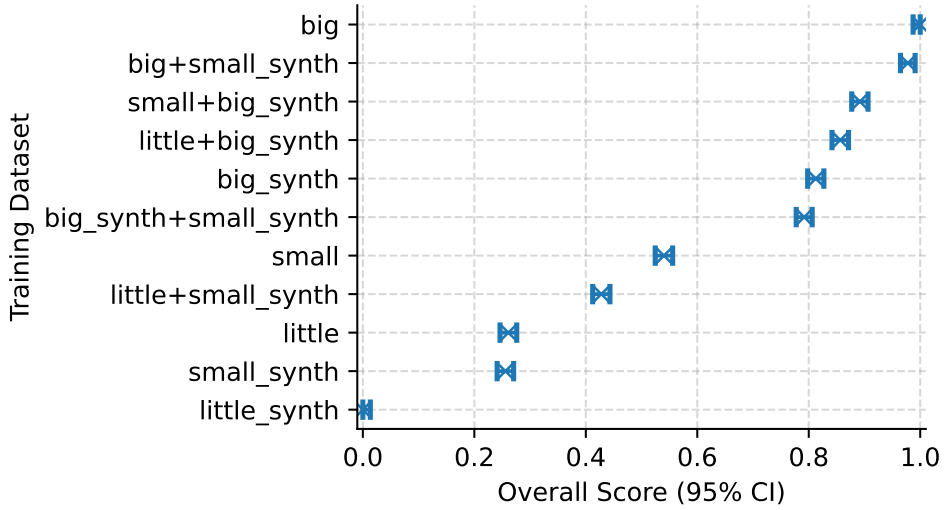


Figure 5.8: Stage 4. Aggregate downstream score for models trained on real, synthetic, and mixed training sets, including Global Generator based augmentation scenarios. Datasets with the suffix `_synth` are purely synthetic and match their real counterparts in size and demographics. Synthetic augmentation markedly improves low resource regimes and approaches the performance of high resource training without sharing patient level data.

Dataset	DRG Classification Accuracy	SOFA Score Prediction R^2	30 day Readmission AUC	ICU Admission AUC	In Hospital Mortality AUC	Overall Score
big	0.740 [0.733, 0.746]	0.582 [0.555, 0.608]	0.771 [0.763, 0.779]	0.913 [0.907, 0.918]	0.916 [0.896, 0.930]	1.000 [0.987, 1.000]
big+small_synth	0.729 [0.723, 0.736]	0.573 [0.548, 0.598]	0.763 [0.755, 0.771]	0.913 [0.908, 0.919]	0.911 [0.893, 0.926]	0.978 [0.965, 0.991]
small+big_synth	0.687 [0.680, 0.695]	0.567 [0.542, 0.590]	0.758 [0.751, 0.766]	0.906 [0.900, 0.912]	0.903 [0.882, 0.919]	0.892 [0.877, 0.906]
little+big_synth	0.669 [0.661, 0.676]	0.566 [0.539, 0.590]	0.756 [0.748, 0.764]	0.907 [0.901, 0.912]	0.898 [0.874, 0.916]	0.857 [0.842, 0.871]
big_synth	0.648 [0.640, 0.655]	0.559 [0.532, 0.584]	0.753 [0.745, 0.761]	0.899 [0.892, 0.905]	0.909 [0.890, 0.924]	0.813 [0.798, 0.827]
big_synth+small_synth	0.638 [0.630, 0.645]	0.556 [0.530, 0.580]	0.746 [0.738, 0.755]	0.898 [0.891, 0.904]	0.880 [0.854, 0.901]	0.792 [0.777, 0.806]
small	0.504 [0.496, 0.512]	0.565 [0.538, 0.590]	0.757 [0.749, 0.766]	0.909 [0.903, 0.915]	0.893 [0.871, 0.909]	0.540 [0.525, 0.556]
little+small_synth	0.450 [0.442, 0.458]	0.550 [0.524, 0.577]	0.741 [0.733, 0.750]	0.898 [0.891, 0.904]	0.894 [0.873, 0.912]	0.428 [0.412, 0.443]
little	0.364 [0.356, 0.371]	0.527 [0.501, 0.552]	0.736 [0.728, 0.744]	0.896 [0.890, 0.902]	0.905 [0.890, 0.918]	0.261 [0.246, 0.276]
small_synth	0.366 [0.358, 0.373]	0.532 [0.502, 0.558]	0.725 [0.716, 0.733]	0.883 [0.876, 0.889]	0.873 [0.850, 0.891]	0.256 [0.241, 0.271]
little_synth	0.240 [0.233, 0.247]	0.485 [0.455, 0.514]	0.710 [0.702, 0.719]	0.857 [0.850, 0.864]	NA	0.000 [0.000, 0.013]

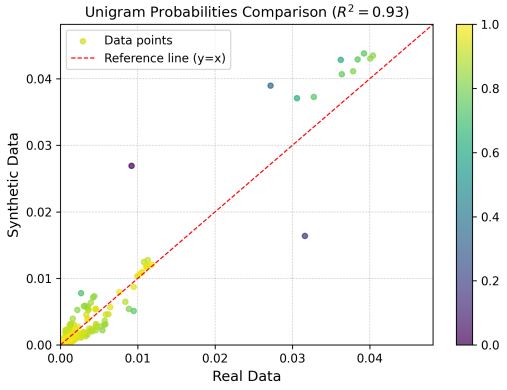
Table 5.2: Results on five downstream tasks for models trained on various combinations of original and synthetic datasets. Names without suffix refer to original data; names ending in `_synth` refer to purely synthetic data; mixed names, for example `big+small_synth`, combine original and synthetic samples. Each cell reports the mean with a 95% confidence interval. NA indicates that results could not be generated due to data scarcity and the fixed model size. The Overall Score column shows the aggregated performance defined in Section 5.2.4.

Temperature	Big		Small		Little	
	Unigram	DimWise	Unigram	DimWise	Unigram	DimWise
0.7	0.930	0.930	0.936	0.937	0.954	0.954
0.9	0.991	0.991	0.992	0.992	0.976	0.976
1.1	0.998	0.998	0.995	0.995	0.955	0.955
1.0	0.999	0.999	0.998	0.998	0.961	0.961

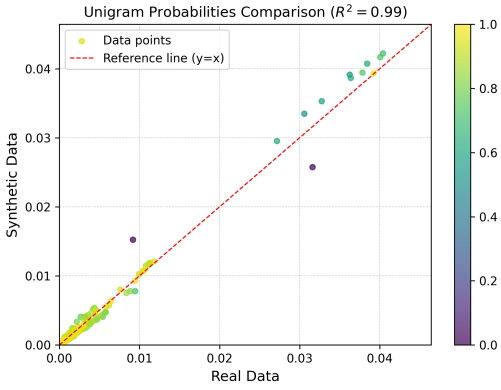
Table 5.3: Fidelity on the timeline dataset across temperatures and data scales.

Temperature	Big		Small		Little	
	Unigram	DimWise	Unigram	DimWise	Unigram	DimWise
0.7	0.959	0.978	0.945	0.983	0.795	0.947
0.9	0.995	0.996	0.995	0.997	0.808	0.934
1.1	0.997	0.999	0.995	0.995	0.954	0.979
1.0	0.999	0.999	0.996	0.996	0.854	0.944

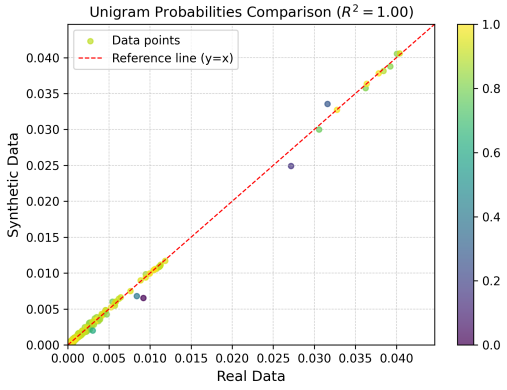
Table 5.4: Fidelity on the readmission dataset across temperatures and data scales.



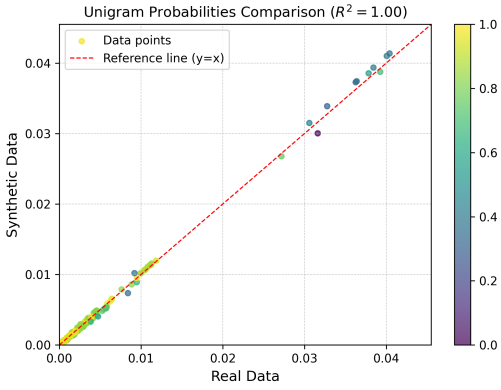
(a) Temperature 0.7



(b) Temperature 0.9



(c) Temperature 1.1



(d) Temperature 1.0

Figure 5.9: Unigram alignment for `big` across temperatures.

6. Synthesis of Contributions

This chapter synthesizes the key contributions of the thesis, integrating insights from three major lines of work: the development of ETHOS as a foundation model for electronic health records, its extension into adaptive and interpretable risk estimation through ARES, and the design of the Federated Timeline Synthesis (FTS) framework for scalable, privacy-preserving deployment across institutions. Together, these contributions articulate a coherent research program that advances the state of the art in healthcare AI, moving toward generalizable, explainable, and interoperable systems capable of supporting real-time decision-making at both individual and population scales.

6.1. ETHOS as a Foundation Model for EHR

ETHOS introduced a unified generative framework for modeling electronic health records (EHRs). The central innovation was the design of *Patient Health Timelines* (PHTs), which tokenized heterogeneous clinical data into a sequential format analogous to natural language. This representation enabled the direct application of transformer architectures—originally developed for text—to healthcare data. By embedding diagnoses, medications, procedures, laboratory results, vital signs, and demographic attributes into a coherent symbolic vocabulary, ETHOS bridged the gap between structured clinical databases and generative modeling frameworks.

Unlike prior approaches, which typically built specialized models for individual tasks, ETHOS was conceived as a general-purpose foundation model. It was trained once on large-scale patient timelines and then applied directly across tasks without further fine-tuning. This eliminated the inefficiencies of bespoke model development and demonstrated that a single representation could serve as the substrate for multiple predictive and generative applications. ETHOS thus established the feasibility of treating healthcare not as a collection of narrowly defined prediction problems, but as a domain where unified foundation models can operate across contexts and institutions.

6.2. Zero-Shot Learning Across Clinical Tasks

A hallmark contribution of ETHOS was the demonstration of *zero-shot learning* in healthcare. ETHOS was evaluated on a diverse set of downstream tasks, including inpatient and ICU mortality, ICU length of stay, 30-day hospital readmission, first-day SOFA scoring, and Diagnosis-Related Group (DRG) classification. Remarkably, the same pretrained model was applied to each of these tasks without task-specific retraining, fine-tuning, or feature engineering.

Performance was not only competitive but often exceeded specialized baselines. ETHOS achieved an AUC of 0.921 for inpatient mortality and 0.927 for ICU mortality, surpassing traditional models such as logistic regression, gradient boosting, and recurrent neural networks. In DRG classification, ETHOS achieved a top-1 accuracy of 84.8%, dramatically outperforming methods reliant on discharge summaries (approximately 52%). Even in tasks considered particularly difficult, such as readmission prediction, ETHOS delivered results comparable to graph neural networks that required extensive preprocessing. These findings underscored the potential of generative

zero-shot inference: a single model, trained once, can generalize across heterogeneous tasks traditionally requiring bespoke architectures.

This paradigm shift—from task-specific to general-purpose modeling—represents one of the most important contributions of ETHOS. It demonstrates that healthcare AI can move beyond fragmented solutions toward foundation models that unify predictive reasoning under a common framework.

6.3. Adaptive, Personalized, and Explainable Risk Estimation

Building on the generative capabilities of ETHOS, the Adaptive Risk Estimation System (ARES) extended the framework into dynamic, patient-specific risk prediction. ARES introduced a novel approach to forecasting outcomes by generating multiple future Patient Health Timelines (fPHTs) and deriving probability distributions from the ensemble of trajectories. This design provides a principled way to quantify risk as a distribution rather than a single point estimate, reflecting the uncertainty inherent in clinical prognostication.

ARES further advanced the interpretability of foundation models in healthcare. Through attribution mechanisms, it identified the specific tokens—such as abnormal laboratory results, acute diagnoses, or medication administrations—that most strongly influenced risk estimates. In addition, ARES generated patient-specific visualizations of evolving risk trajectories, enabling clinicians to see not only a risk score but also the historical events and future scenarios shaping that estimate. This dual focus on personalization and transparency addresses two of the most significant barriers to clinical adoption: the need for individualized predictions and the need for models to explain themselves in clinically meaningful terms.

ARES thus demonstrated that foundation models can be more than black-box predictors. With carefully designed extensions, they can provide adaptive, interpretable decision-support systems that integrate seamlessly with clinician reasoning.

6.4. Federated Scalability and Privacy-Preserving Deployment

While ETHOS and ARES proved effective within single-institution datasets, healthcare AI must ultimately scale across hospitals, regions, and countries. The Federated Timeline Synthesis (FTS) framework was developed to meet this challenge. FTS enables multi-institutional collaboration without requiring the centralization of sensitive patient records. Instead, each institution generates synthetic patient timelines that preserve the statistical and temporal structure of local data while ensuring privacy. These synthetic timelines can then be shared for federated training, providing the benefits of large-scale data without the risks of data leakage.

FTS addressed two critical challenges simultaneously. First, it provided strong privacy protections, allowing institutions to collaborate without compromising patient confidentiality or violating data-sharing regulations. Second, it established a pathway to interoperability by standardizing data into a common tokenized format. By ensuring that synthetic PHTs retained both clinical semantics and temporal coherence, FTS allowed ETHOS to be trained on globally diverse datasets while mitigating site-specific biases.

This contribution is particularly significant for scaling foundation models in healthcare. Whereas most prior work has been confined to single datasets, FTS enables ETHOS to become a truly global model, incorporating diverse populations while respecting institutional and regulatory boundaries.

6.5. Comparison with Other Foundation Model Approaches

ETHOS and its extensions can be situated within the broader ecosystem of healthcare foundation models, which can be divided into two primary categories. The first includes Clinical Language Models (CLaMs), such as BioBERT, ClinicalBERT, and GatorTron, which are trained predominantly on unstructured biomedical and clinical text. These models excel in tasks such as question answering, summarization, and knowledge extraction from narrative data. The second includes Foundation Models for Electronic Medical Records (FEMRs), which focus on structured EHR data and typically produce embeddings for downstream classifiers to predict outcomes or phenotypes.

ETHOS diverges from both paradigms in critical ways. Unlike CLaMs, ETHOS does not restrict itself to text but instead tokenizes structured events, enabling the direct modeling of longitudinal patient trajectories. Unlike FEMRs, which are primarily embedding generators for classifiers, ETHOS is generative: it simulates future timelines, enabling zero-shot inference and scenario analysis. In combining these two qualities—structured event modeling and generative forecasting—ETHOS establishes a new class of foundation models uniquely suited to healthcare.

6.6. A Coherent Research Trajectory

Taken together, the contributions of ETHOS, ARES, and FTS represent a coherent progression. ETHOS demonstrated the feasibility of treating patient records as tokenized timelines and applying transformers to simulate and predict health trajectories. ARES extended this framework into adaptive, personalized, and explainable risk estimation, addressing critical needs for trust and interpretability in clinical settings. FTS expanded the scope of deployment to multi-institutional environments, providing a scalable and privacy-preserving strategy for global collaboration.

This progression reflects a broader vision: the creation of foundation models for healthcare that are not only accurate but also generalizable, interpretable, and deployable across institutional boundaries. ETHOS laid the foundation, ARES made it actionable at the bedside, and FTS ensured it could scale across the healthcare ecosystem. Together, they chart a path toward foundation models that unify prediction, explanation, and simulation, paving the way for the eventual realization of patient digital twins and large-scale, equitable healthcare AI.

7. Future Directions

The development of ETHOS, ARES, and the Federated Timeline Synthesis (FTS) framework represents a major step forward in healthcare AI, demonstrating that transformer-based foundation models can operate on structured patient timelines, perform zero-shot prediction, and scale across institutions while preserving privacy. These contributions provide both proof-of-concept and a foundation for further innovation. At the same time, they open up a wide range of avenues for future research and development. In this chapter, we articulate key directions that can expand the clinical relevance, technical robustness, and societal value of foundation models for healthcare. Each direction is presented as a domain of focus, reflecting the multidimensional challenges that must be addressed to transform ETHOS from a research prototype into a comprehensive, globally deployable system.

7.1. Integration of Multimodal Data

A critical next step for ETHOS is the integration of multimodal data streams beyond the structured events available in current EHR systems. While the ETHOS framework effectively captures diagnoses, procedures, medications, vital signs, and laboratory values, real-world patient care extends far beyond these coded fields. Several complementary modalities hold promise for enhancing both the breadth and depth of patient representation:

- **Narrative clinical notes.** Physicians' free-text documentation encodes reasoning, context, and subtlety not captured in structured fields. Progress notes, discharge summaries, and consult letters contain information about social circumstances, differential diagnoses, and subjective impressions, all of which influence patient trajectories.
- **Medical imaging.** Radiology scans (CT, MRI, X-ray) and pathology slides provide visual confirmation of disease states. Integrating imaging embeddings with structured tokens would allow ETHOS to correlate coded diagnoses with phenotypic evidence of disease progression.
- **Omics data.** Genomic, transcriptomic, and proteomic profiles represent biological predispositions and drug sensitivities, providing individualized context for clinical decision-making. Their integration would enable ETHOS to generate forecasts at the intersection of genotype and phenotype.
- **Continuous monitoring and wearables.** Bedside monitors in the ICU and consumer-grade devices outside the hospital produce rich, high-frequency physiological time series. These signals offer early-warning indicators of deterioration and can anchor predictive models to the real-time patient state.
- **Environmental and behavioral data.** Social determinants of health, mobility data, and lifestyle measures provide further dimensions that influence outcomes and must eventually be represented in holistic patient timelines.

Technically, multimodal fusion may be achieved through cross-attention modules that allow ETHOS to jointly attend to structured tokens, textual embeddings, and image or waveform representations. The challenge is to balance

modality-specific encoding with a unified representation that preserves interpretability. Success in this area could transform ETHOS from a timeline model into a comprehensive "digital health atlas," capturing the full spectrum of information relevant to patient care.

7.2. Real-Time Clinical Deployment and Usability

To translate ETHOS into clinical impact, research must focus on deployment within real-world workflows. Even the most accurate predictive model will fail to improve outcomes if it is not accessible, interpretable, and actionable for clinicians. Several areas demand attention:

- **Latency and efficiency.** ETHOS has already demonstrated inference times of 1–30 seconds per patient on modern GPUs, which is sufficient for bedside or ward-level use. However, deployment at scale requires optimization for hospital hardware constraints, including CPUs and low-resource environments.
- **User interface design.** Predictions must be displayed in ways that align with clinical cognition. Possible approaches include dashboards showing evolving patient trajectories, risk stratification summaries, and "traffic-light" alerts for deterioration. Care must be taken to avoid alert fatigue by ensuring that outputs are contextually relevant and clinically justified.
- **Decision support integration.** ETHOS outputs should be embedded directly into EHR systems to support decision-making in real time, rather than requiring clinicians to consult external platforms. APIs and middleware may play a role here.
- **Scenario exploration.** Clinicians must be able to query ETHOS interactively, exploring "what-if" scenarios to understand how different treatment strategies could alter predicted outcomes.

Beyond technical challenges, deployment requires sociotechnical alignment: trust-building with clinicians, the explainability of model outputs, and iterative co-design with healthcare professionals. Future work must explore how ETHOS predictions influence clinical workflows, patient outcomes, and healthcare resource use in practice.

7.3. Fairness, Bias Mitigation, and Generalizability

AI systems in healthcare must not only be accurate but also equitable. Bias in predictions can reinforce existing disparities, leading to harmful outcomes. ETHOS and ARES demonstrated promising subgroup performance on race and gender within MIMIC-IV, but future work must broaden and deepen these analyzes:

- **Cross-institutional generalizability.** Validation must extend to diverse hospitals, regions, and countries. Models trained on U.S. academic medical centers may not generalize to rural clinics or international health systems with different coding practices.
- **Underrepresented groups.** ETHOS must be systematically tested on populations that are underrepresented in the training data, such as rare disease cohorts, children, or individuals with atypical care trajectories.
- **Bias in generative simulations.** Because ETHOS generates probabilistic future patient health timelines (fPHTs), any bias in the training data could manifest as skewed distributions of predicted outcomes. Fairness-aware training strategies, bias correction layers, and post-hoc calibration must be explored.
- **Evaluation frameworks.** Beyond aggregate metrics such as AUC, fairness must be evaluated using subgroup calibration, equal opportunity, and other fairness-specific metrics.

Addressing these concerns will require both methodological innovations and institutional collaborations. Techniques such as adversarial debiasing, domain adaptation, and federated fairness audits may become central to ensuring that ETHOS predictions are equitable across patient populations.

7.4. Federated and Privacy-Preserving Learning at Scale

The Federated Timeline Synthesis (FTS) framework provides a blueprint for training foundation models across distributed health systems without centralizing sensitive patient data. Building on this foundation, several directions are apparent:

– **Scalability.** Extending federated training to networks of hundreds of hospitals, each with millions of patient records, will require algorithmic advances in communication efficiency, synchronization strategies, and robustness to non-IID (non-identically distributed) data.

– **Privacy guarantees.** Differential privacy, secure multiparty computation, and homomorphic encryption may be layered onto FTS to provide formal guarantees that no patient-level information is leaked.

– **Cross-border collaborations.** Federated learning could allow ETHOS to be trained on global datasets, harmonizing insights across countries while respecting national data protection laws, such as GDPR.

– **Benchmarking.** Standardized protocols for evaluating federated foundation models will be essential for comparing methods and establishing trust among institutions.

By advancing federated learning, ETHOS can move from a single-institution model to a globally trained system, incorporating diverse patient populations while preserving privacy and sovereignty.

7.5. Interpretability and Human-Centered AI

For adoption in clinical settings, ETHOS must not be a black box. Interpretability is central to building clinician trust and ensuring responsible use. Several directions merit exploration:

– **Attention visualization.** Mapping which tokens ETHOS attends to when predicting outcomes could highlight clinically salient events, aiding explanation and validation.

– **Counterfactual reasoning.** ETHOS could simulate how small changes in patient history (e.g., administering or withholding a drug) alter predicted outcomes, providing causal insights.

– **Hierarchical token semantics.** The multi-token representation of codes (e.g., ICD, ATC) lends itself to an interpretable decomposition of predictions into contributions from different levels of specificity.

– **Embedding analysis.** Learned token embeddings, which already show clustering by diagnostic group and ordinal progression for quantiles, could be further developed into tools for discovering latent structure in clinical data.

These approaches could transform ETHOS from a predictive tool into a partner in clinical reasoning, offering explanations that resonate with both data-driven and domain-driven perspectives.

7.6. Towards Patient Digital Twins

Perhaps the most ambitious vision is for ETHOS to serve as the computational backbone for patient digital twins: individualized, continuously updated models that simulate a person's future health trajectory and responses to interventions. ETHOS already provides many of the foundational capabilities required, including generative timeline modeling, risk estimation, and counterfactual scenario analysis. Future development toward digital twins will involve:

- **Continuous updating.** Real-time integration of new data as it becomes available, from EHR updates to wearable devices, ensuring the digital twin evolves alongside the patient.
- **Therapeutic simulation:** Modeling alternative treatment strategies and projecting their outcomes, thereby enabling personalized medicine and shared decision-making between clinicians and patients.
- **Population-level scaling.** Aggregating digital twins to simulate interventions at the population level, supporting health policy design and evaluation.
- **Economic integration.** Extending simulations to predict not only clinical outcomes but also costs, resource use, and quality-adjusted life years (QALYs), aligning AI-driven care planning with value-based healthcare.

The realization of patient digital twins would transform ETHOS from a predictive model into a dynamic simulation engine, capable of guiding prevention, acute care, chronic disease management, and policy decisions throughout the lifespan.

7.7. Ethical, Regulatory, and Societal Considerations

Finally, future research must address the broader societal context in which ETHOS operates. Several domains require sustained attention:

- **Ethical governance.** Frameworks for consent, accountability, and oversight must be developed to ensure that foundation models are deployed responsibly.
- **Regulatory pathways.** Clear guidelines for validation, approval, and post-deployment monitoring will be necessary for adoption in clinical care.
- **Societal impact.** The deployment of ETHOS at scale could reshape workflows, shift responsibilities, and even influence health economics. Anticipating these impacts is critical to avoid unintended consequences.
- **Education and training.** Clinicians and patients alike must be equipped to understand, interpret, and act upon model outputs in ways that enhance rather than undermine clinical judgment and patient autonomy.

By embedding ethical and regulatory considerations into technical development, ETHOS can evolve into a trustworthy foundation for healthcare AI.

8. Conclusion

8.1. Summary of Findings

This dissertation set out to test the hypothesis that tokenized patient health timelines, modeled through generative transformer architectures, can serve as a universal representation of electronic health records (EHRs). The research demonstrated that this framework is feasible and impactful through three key systems: ETHOS, a foundation model for zero-shot health trajectory prediction; ARES, an adaptive risk estimation system built upon generative forecasting; and FTS, a federated synthesis methodology enabling scalable and privacy-preserving cross-institutional training. These contributions collectively prove the central hypothesis and provide a coherent pathway toward trustworthy and general-purpose clinical AI.

8.2. Validation of the Hypothesis

Across ETHOS, ARES, and FTS, the central claims of the thesis have been rigorously validated:

– **Zero-shot generalization:** ETHOS achieved competitive or superior performance compared to task-specific models across diverse prediction tasks without fine-tuning.

– **Adaptive inference:** ARES delivered interpretable, personalized risk estimates through simulations of future health trajectories.

– **Federated deployment:** FTS enabled cross-institutional collaboration without sharing raw patient data, addressing one of the most critical barriers to healthcare AI.

Together, these results confirm that a single generative modeling framework can unify prediction, simulation, explainability, and federated training at scale.

8.3. Adoption, Strengths, and Future Impact

Since its introduction less than a year ago, ETHOS has achieved rapid and wide adoption—already cited 32 times—signaling a transformative shift in how patient timelines are modeled and evaluated. Beyond academic uptake, leading institutions have built directly on ETHOS: Epic Systems, Microsoft Research, and Yale University developed the Cosmos Medical Event Transformer (CoMET) [52], scaling the paradigm to over 118,million patients and 115,billion medical events and validating reliable scaling laws across diverse outcomes; in parallel, Microsoft Research has systematically examined scaling laws for EHR foundation models [62]. Interest from Google, Verily, NVIDIA, health-tech startups, and insurance providers further underscores ETHOS as a practical foundation for clinical prediction, operational optimization, and personalized modeling. This impact is enabled

by core innovations introduced in this thesis: the first EHR foundation model capable of zero-shot trajectory prediction across heterogeneous tasks; adaptive, explainable inference via generative simulation of patient futures; privacy-preserving federated synthesis to scale models across institutions; and a strong commitment to open science—releasing code, models, and the MEDS to catalyze reproducibility and community progress. Looking ahead, extending ETHOS to multimodal streams (notes, imaging, genomics, wearables) will power patient digital twins that support individualized, lifelong simulations of outcomes, interventions, and costs; bias and fairness analysis will promote equitable performance across populations; and real-time inference latencies of 1–30s per patient make clinical integration feasible today. Taken together, these adoption signals and technical strengths demonstrate both scientific originality and immediate utility, with **potential to reshape how medicine is practiced** by shifting care from reactive to predictive, proactive, and precisely targeted at scale.

It is a privilege to have contributed to this paradigm shift, and the results presented here mark a decisive step toward a future where foundation models play a central role in medicine worldwide.

Bibliography

[1] MEDS-DEV: Establishing reproducibility and comparability in health AI. `https://github.com/mmcdermott/MEDS-DEV`.

[2] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.

[3] David W Bates, Hsiang-Yin Cheng, NT Cheung, Rita Jew, Fraz Mir, Robyn Tamblyn, and Yu-Chuan Li. 'Improving smart medication management': an online expert discussion. *BMJ Health & Care Informatics*, 29(1):e100540, April 2022.

[4] Monik Raj Behera, Sudhir Upadhyay, Suresh Shetty, Sudha Priyadarshini, Palka Patel, and Ker Farn Lee. Fedsyn: Synthetic data generation using federated learning. *arXiv preprint arXiv:2203.05931*, 2022.

[5] Alban Bornet, Dimitrios Proios, Anthony Yazdani, Fernando Jaume-Santero, Guy Haller, Edward Choi, and Douglas Teodoro. Comparing neural language models for medical concept representation and patient trajectory prediction, June 2023. Pages: 2023.06.01.23290824.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. arXiv:2005.14165 [cs].

[7] Junde Chen, Trudi Di Qi, Jacqueline Vu, and Yuxin Wen. A deep learning approach for inpatient length of stay and mortality prediction. *Journal of Biomedical Informatics*, 147:104526, November 2023.

[8] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, August 2016. ACM.

[9] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[10] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.

[11] Geunho Choi, Won Chul Cha, Se Uk Lee, and Soo-Yong Shin. Survey of medical applications of federated learning. *Healthcare Informatics Research*, 30(1):3–15, 2024.

[12] Matthew M Churpek, Trevor C Yuen, Seo Young Park, David O Meltzer, Jesse B Hall, and Dana P Edelson. Derivation of a cardiac arrest prediction model using ward vital signs. *Crit. Care Med.*, 40(7):2102–2108, July 2012.

[13] David R Eitel, Debbie A Travers, Alexander M Rosenau, Nicki Gilboy, and Richard C Wuerz. The emergency severity index triage algorithm version 2 is reliable and valid. *Acad. Emerg. Med.*, 10(10):1070–1080, October 2003.

[14] Micah Hartman, Anne B Martin, Lekha Whittle, Aaron Catlin, and National Health Expenditure Accounts Team. National health care spending in 2022: Growth similar to prepandemic rates. *Health Aff.*, 43(1):6–17, January 2024.

[15] S Hochreiter and J Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[16] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, November 2020. arXiv:1904.05342 [cs].

[17] Shaoxiong Ji, Yue Tan, Teemu Saravirta, Zhiqin Yang, Yixin Liu, Lauri Vasankari, Shirui Pan, Guodong Long, and Anwar Walid. Emerging trends in federated learning: From model fusion to federated x learning. *International Journal of Machine Learning and Cybernetics*, 15(9):3769–3790, 2024.

[18] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV, 2023.

[19] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. Publisher: Nature Publishing Group.

[20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[21] Ivana Kolic, Smiley Crane, Suzanne McCartney, Zane Perkins, and Alex Taylor. Factors affecting response to national early warning score (news). *Resuscitation*, 90:85–90, May 2015.

[22] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. Foresight–a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, April 2024. Publisher: Elsevier.

[23] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020.

[24] Xia Li, Meng Yu, Rui Yang, Wei Guan, and Xiangyang Jiang. Fed-bert: A federated pre-training framework for clinical nlp. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13211–13219, 2021.

[25] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, February 2023.

[26] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1):7155, April 2020. Publisher: Nature Publishing Group.

[27] Xiao Ling, Tim Menzies, Christopher Hazard, Jack Shu, and Jacob Beel. Trading off scalability, privacy, and performance in data synthesis. *IEEE Access*, 12:26642–26654, 2024.

[28] Claire Little, Mark Elliot, and Richard Allmendinger. Federated learning for generating synthetic data: a scoping review. *International Journal of Population Data Science*, 8(1):2158, 2023.

[29] Bingyan Liu, Nuoyan Lv, Yuanchun Guo, and Yawen Li. Recent advances on federated learning: A systematic survey. *Neurocomputing*, page 128019, 2024.

[30] Matthew B. A. McDermott, Bret Nestor, Peniel Argaw, and Isaac Kohane. Event Stream GPT: A Data Preprocessing and Modeling Library for Generative, Pre-trained Transformers over Continuous-time Sequences of Complex Events, June 2023. arXiv:2306.11547 [cs].

[31] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *arXiv [cs.LG]*, 54:1273–1282, February 2016.

[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. *J. Intern. Med.*, 255(5):579–587, May 2004.

[33] Nassim Oufattole, Teya Bergamaschi, Aleksia Kolo, Hyewon Jeong, Hanna Gaggin, Collin M Stultz, and Matthew B A McDermott. MEDS-tab: Automated tabularization and baseline methods for MEDS datasets. *arXiv [cs.LG]*, October 2024.

[34] Xiaobin Pan, Jinbao Xie, Lihui Zhang, Xincai Wang, Shujuan Zhang, Yingfeng Zhuang, Xingsheng Lin, Songjing Shi, Songchang Shi, and Wei Lin. Evaluate prognostic accuracy of SOFA component score for mortality among adults with sepsis by machine learning method. *BMC Infectious Diseases*, 23(1):76, February 2023.

[35] Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, M Spotnitz, Ruijun Chen, A Perotte, and K Natarajan. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. 158:239–260, November 2021.

[36] Ke Pang, Liang Li, Wen Ouyang, Xing Liu, and Yongzhong Tang. Establishment of ICU Mortality Risk Prediction Models with Machine Learning Algorithm Using MIMIC-IV Database. *Diagnostics*, 12(5):1068, April 2022.

[37] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[38] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):1–13, May 2021. Publisher: Nature Publishing Group.

[39] Pawel Renc, Michal K. Grzeszczyk, Linglong Qian, Nassim Oufattole, Jeff Rasley, and Arkadiusz Sitek. Federated timeline synthesis: Scalable and private methodology for model training and deployment, 2025.

[40] Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ Digit. Med.*, 7(1):256, September 2024.

[41] Eric C. Schneider and Reginald D. Williams II. Mirror, Mirror 2021: Reflecting Poorly, August 2021.

[42] Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. 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. *Resuscitation*, 84(4):465–470, April 2013.

[43] Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. *arXiv [cs.LG]*, January 2023.

[44] C P Subbe, M Kruger, P Rutherford, and L Gemmel. Validation of a modified early warning score in medical admissions. *QJM*, 94(10):521–526, October 2001.

[45] Siyi Tang, Amara Tariq, Jared A. Dunnmon, Umesh Sharma, Praneetha Elugunti, Daniel L. Rubin, Bhavik N. Patel, and Imon Banerjee. Predicting 30-Day All-Cause Hospital Readmission Using Multimodal Spatiotemporal Graph Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, 27(4):2071–2082, April 2023. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[46] Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature Communications*, 14(1):5305, August 2023. Publisher: Nature Publishing Group.

[47] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, August 2023. Publisher: Nature Publishing Group.

[48] Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, Eric J. G. Sijbrands, Erwin J. O. Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, Gilles Clermont, Mihaela Van Der Schaar, Ari Ercole, Armand R. J. Girbes, and Paul W. G. Elbers. Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: The amsterdam university medical centers database (amsterdamumcdb) example∗. *Critical Care Medicine*, 49(6):e563–e577, june 2021.

[49] Amirsina Torfi, Edward A Fox, and Chandan K Reddy. Differentially private synthetic medical data generation using convolutional gans. *Information Sciences*, 586:485–500, 2022.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30 of *NIPS'17*, pages 6000–6010, Red Hook, NY, USA, December 2017. Curran Associates Inc. arXiv:1706.03762 [cs].

[51] Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine*, 7(1):1–9, January 2024. Publisher: Nature Publishing Group.

[52] Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale. (arXiv:2508.12104), August 2025. arXiv:2508.12104 [cs].

[53] John Weldon, Tomas Ward, and Eoin Brophy. Generation of synthetic electronic health records using a federated gan. *arXiv preprint arXiv:2109.02543*, 2021.

[54] Bryan Williams. The national early warning score: from concept to NHS implementation. *Clin. Med.*, 22(6):499–505, November 2022.

[55] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models, December 2023. arXiv:2307.02028 [cs].

[56] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):1–10, July 2023. Publisher: Nature Publishing Group.

[57] Feng Xie, Bibhas Chakraborty, Marcus Eng Hock Ong, Benjamin Alan Goldstein, Nan Liu, et al. Autoscore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR medical informatics*, 8(10):e21798, 2020.

[58] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *npj Digital Medicine*, 5(1):1–9, December 2022. Publisher: Nature Publishing Group.

[59] Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14(1):7857, November 2023. Publisher: Nature Publishing Group.

[60] Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, et al. Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ digital medicine*, 6(1):141, 2023.

[61] Betul Yurdem, Murat Kuzlu, Mehmet Kemal Gullu, Ferhat Ozgur Catak, and Maliha Tabassum. Federated learning: Overview, strategies, applications, tools and future directions. *Heliyon*, 2024.

[62] Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. Exploring scaling laws for ehr foundation models. (arXiv:2505.22964), May 2025. arXiv:2505.22964 [cs].

[63] Tongyan Zhang, Yazhu Hou, Yan Li, Xin Yang, Shengyuan Zhou, Guoxian Lu, Pengyun Shen, and Xiumei Gao. National early warning score (NEWS) system for improving response time in an acute care setting: A retrospective study. *Am. J. Emerg. Med.*, 87:209–212, January 2025.

[64] Guanglin Zhou and Sebastiano Barbieri. Generating clinically realistic ehr data via a hierarchy-and semantics-guided transformer. *arXiv preprint arXiv:2502.20719*, 2025.