HIERARCHICAL RELATIONAL TRAJECTORY MODELING FOR LONGITUDINAL CLINICAL PROGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prognosis from longitudinal electronic health records requires integrating evidence that is generated across hospitalizations and days, yet most existing models do not align their modeling of such records with this process, so a prediction cannot be traced back to the evidence it rests on, and whether a model actually relies on the trajectory cannot be checked. To address this mismatch, we propose **HiReT**, a hierarchical relational trajectory model that not only reconstructs patient records into a `visit-day-node` hierarchy, but also explicitly models the dynamic dependencies among heterogeneous evidence. **HiReT** expresses the dependencies among evidence types with a typed relation map within each visit, and uses hierarchical aggregation to follow the clinical data flow, combining days into visits and visits into a patient representation. Experiments on a natural-distribution six-class mortality benchmark from MIMIC-IV show that **HiReT** achieves relative improvements of 19.95%, 27.81%, and 47.39% in Balanced Accuracy, Macro-F1, and quadratic weighted kappa (QWK), respectively, over the strongest baseline, and is the best performer on all five metrics of a prolonged length-of-stay (LOS) benchmark. Furthermore, ablations over the model's structure and evidence types, together with pathway-level interventions, show that the model's predictions depend on specific clinical evidence and on the order of the trajectory, and that the contribution of individual evidence types and pathways can be probed directly. Overall, these results indicate that aligning a model with how clinical evidence is generated improves prognosis accuracy and makes the evidence its predictions rest on more directly probeable.

1 INTRODUCTION

Clinical prognosis prediction aims to estimate a patient's future clinical outcomes, such as mortality, disease progression, or post-discharge survival. Accurate prognosis models are clinically valuable for risk stratification, treatment planning, and resource allocation (Tu et al., 2024). Prior work has studied such prediction on longitudinal electronic health records (EHRs), in settings that range from standardized outcome benchmarks to instruction-based temporal modeling (Lovon-Melgarejo et al., 2024; Grundmann et al., 2026; Cui et al., 2025). In this work, we study this task on longitudinal EHRs, which contain clinical evidence such as demographics, laboratory tests, medications, procedures, careflow (transfers and service changes), clinical notes, and reports (Johnson et al., 2023; Elsharief et al., 2025). Such records span multiple temporal scales, from hospitalizations to the days within each one, and the evidence recorded at those scales is heterogeneous in type, dispersed in time, and dependent across types; accurate prognosis therefore depends on how such evidence is integrated.

These properties make prognosis prediction challenging. Patient histories in EHRs often extend over long temporal horizons, requiring models to integrate information across repeated visits and hospitalizations (Cui et al., 2025; Lovon-Melgarejo et al., 2024). In addition, structured clinical records are highly heterogeneous, comprising diverse clinical information together with repeated measurements over time (Elsharief et al., 2025; Yao et al., 2024). As illustrated in the raw-record panel of Figure 1, clinically relevant evidence may be widely dispersed across long time spans, while dependencies among different evidence types evolve within each hospitalization. When the record is serialized into a flat event stream, these boundaries and dependencies are erased, and the model can only recover them implicitly from a mixed sequence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing approaches to longitudinal prognosis prediction differ in how they treat this structure. Some represent patient history as a linearized sequence. RETAIN (Choi et al., 2016) and HiTANet (Luo et al., 2020) use attention to model visit-level dynamics, while BEHRT (Li et al., 2020), Med-BERT (Rasmy et al., 2021), and CEHR-BERT (Pang et al., 2021) use Transformers and pretraining to capture dependencies across visits or events; foundation-model-based approaches similarly process serialized records (Lovon-Melgarejo et al., 2024; Sellergren et al., 2026). These methods have steadily improved long-range dependency modeling, but the boundaries between hospitalizations and days, and the dependencies among evidence types, remain implicit. Others introduce explicit hierarchy or relations. Hi-BEHRT organizes long EHR sequences with a hierarchical Transformer (Li et al., 2022); CACHE models higher-order code relations through hypergraph-based factual and counterfactual reasoning (Xu et al., 2022); and CRISP incorporates structured prior knowledge for ICU mortality prediction (Wang et al., 2025). These structures are used mainly for long-sequence context compression, code-level relational reasoning, or task-specific guidance, and rarely align, from the perspective of the clinical trajectory, the relations among different clinical evidence with the temporal progression of the trajectory. Whether patient history is modeled as a linearized sequence or with explicit hierarchy and relations, these approaches share one limitation: their modeling is not aligned with the process in which clinical evidence is generated. As a result, the basis of their predictions cannot be traced back to the clinical evidence those predictions rest on, which in turn prevents effective downstream clinical analysis.

To address this misalignment, we propose HiReT, a hierarchical relational trajectory model for clinical prognosis prediction. Concretely, HiReT reconstructs the longitudinal clinical trajectory as a `visit-day-node` hierarchy: the visit and day levels capture temporal progression across and within hospitalizations, and within each visit a typed relation map captures dependencies among evidence types, with the visit background conditioning the dynamic evidence and the dynamic evidence types exchanging information with one another. DayFuse then fuses each day's nodes into a day representation, and hierarchical aggregation combines days into visits and visits into a patient representation for prediction. Figure 1 shows how this data organization defines prediction examples for the two tasks. On a natural distribution six-class mortality benchmark constructed from MIMIC-IV (Johnson et al., 2023), HiReT improves Balanced Accuracy, Macro-F1, and quadratic weighted kappa (QWK) by 4.49, 6.13, and 14.10 points over the strongest baseline (19.95%, 27.81%, and 47.39% relative), and on prolonged length-of-stay (LOS) prediction it is the best performer on all five metrics. We further conduct structural and evidence-type ablations and perturb the input structure by removing or reversing historical visits. The results show that the model leverages both clinical evidence from different sources and the temporal hierarchy of the trajectory: the patient's basic information and admission context are most decisive for prognosis, while treatment and measurement records play distinct roles in separating death from survival and in localizing risk along the outcome timeline.

In summary, our contributions are threefold. (1) We introduce HiReT, a hierarchical relational trajectory model that resolves the mismatch between how existing methods model a patient's record and how clinical evidence is generated and documented along a patient's trajectory. (2) We realize this design as a `visit-day-node` hierarchy that builds a patient representation progressively: a typed relation map carries the visit background to the dynamic evidence types and exchanges information among them, DayFuse fuses the nodes of each day, and hierarchical aggregation combines days into visits and visits into a patient representation. (3) We evaluate HiReT on a six-class mortality benchmark from MIMIC-IV, where it improves class-balanced and ordinal metrics over the strongest baseline, and on the LOS benchmark built from the same dataset, where it achieves the best result on every metric. Ablations over the model's structural components and evidence types show that HiReT leverages clinical evidence from different sources and the trajectory's temporal hierarchy across hospitalizations and days.

2 Related Work

Linearized Sequence Modeling. A dominant line formulates longitudinal electronic health records (EHRs) as one serialized sequence and models temporal dependence across visits. RETAIN identifies influential visits with reverse-time attention (Choi et al., 2016), and HiTANet adds hierarchical time-aware attention for irregular intervals (Luo et al., 2020); related models refine visit dependence, adaptive temporal representations, and missing-value han-

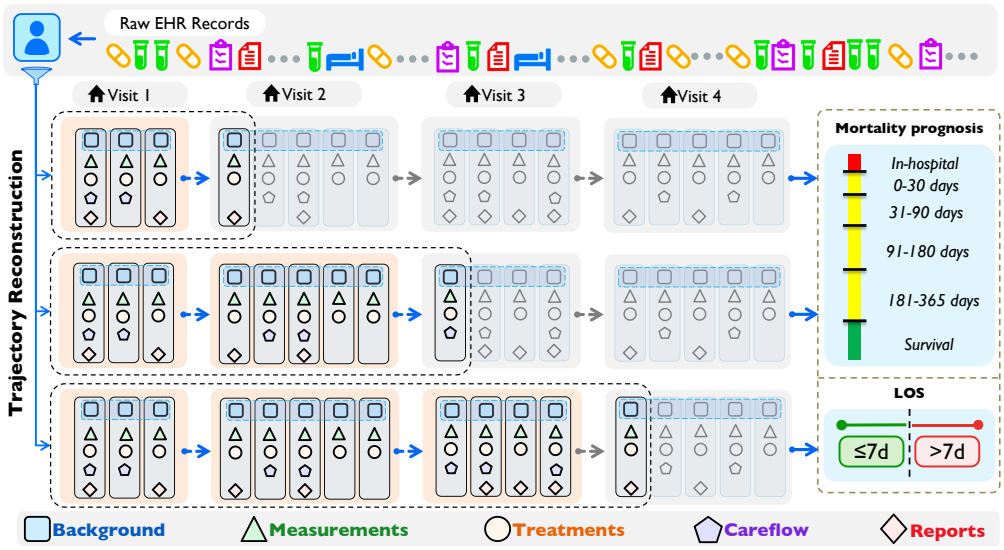

Figure 1: Overview of index-specific trajectory reconstruction and prediction. Raw EHR records are organized into visit background and daily typed evidence. Each rectangle within a visit represents one calendar day (24 hours). The three rows correspond to index admissions 2, 3, and 4 of the same patient, each combining completed historical visits with current-visit evidence available by the prediction landmark. Colored elements indicate available inputs; gray elements indicate evidence unavailable at prediction time. A visit has a single patient background shared by all of its days, so the background drawn on every day of a visit is the same node. Each index example supports six-class mortality prognosis and binary total length-of-stay (LOS) prediction (≤ 7 versus > 7 days).

dling (Ma et al., 2017; 2020; Che et al., 2018). BEHRT adapts Transformers to longitudinal EHR modeling (Li et al., 2020), Med-BERT adds large-scale self-supervised pretraining (Rasmy et al., 2021), and CEHR-BERT incorporates temporal information into contextual representations (Pang et al., 2021); later variants extend pretraining objectives and architectures (Yang et al., 2023; Kraljevic et al., 2024). Foundation models carry this line to instruction tuning and clinical reasoning: TIMER studies temporal instruction modeling (Cui et al., 2025), recent benchmarks evaluate medical and general-purpose LLMs on serialized records (Lovon-Melgarejo et al., 2024; Grundmann et al., 2026), and representative models include HuatuoGPT (Chen et al., 2024), MedGemma (Sellergren et al., 2025; 2026), Lingshu (Xu et al., 2025), and Qwen3 (Yang et al., 2025). All of these methods still consume one serialized sequence, so the boundaries between hospitalizations and days remain implicit.

Explicit Hierarchy and Relations. A second line builds explicit structure into EHR modeling. Hierarchical attention networks established multi-level representations for document classification (Yang et al., 2016), and hierarchical Transformers extended them to long clinical documents (Si & Roberts, 2021); Hi-BEHRT applies hierarchical Transformers to multimodal longitudinal EHRs (Li et al., 2022), and related methods explore multi-level representations for long patient histories (Wang et al., 2023; Du et al., 2022). A complementary direction models relations among clinical concepts: CACHE captures higher-order code dependencies through hypergraph-based counterfactual reasoning (Xu et al., 2022), CRISP injects structured clinical priors for ICU mortality prediction (Wang et al., 2025), and graph-based or knowledge-guided methods likewise exploit explicit concept relations (Jiang et al., 2024; Carvalho et al., 2023). These structures add stronger inductive biases than serialized models, but they are not derived from how clinical evidence is generated along a patient path.

3 METHOD

3.1 PROBLEM FORMULATION

Prediction setting. For patient p, let V_N be an index hospitalization and τ the prediction landmark. HiReT observes all completed historical visits and the evidence available from the index visit by τ:

$$\mathcal{X}_p^{(N)} = (V_1, \ldots, V_{N-1}, V_N^{\leq \tau}). \tag{1}$$

In our benchmarks, τ is 24 hours after index admission. A patient may contribute examples with different index visits; every example contains only information available at its own landmark. We write visit and day indices as i and j and suppress the patient index, which returns as the batch index n in the loss; the number of visits and the number of days vary across examples.

Prediction tasks. The primary task predicts six ordered outcomes: death during the index hospitalization; death within 0–30, 31–90, 91–180, or 181–365 days after discharge; and survival through 365 days after index discharge, supported by follow-up evidence. Survival is defined at this specified horizon: both patients known to die after day 365 and those observed alive through the horizon belong to this class, so the task separates death from survival and localizes death to a window. Examples without sufficient follow-up to establish survival are excluded (Appendix A.1). We also study whether the total index-hospital length of stay (LOS) exceeds seven days. The two task-specific models use the same HiReT architecture with different prediction heads and are trained separately.

3.2 VISIT–DAY–NODE TRAJECTORY RECONSTRUCTION

Hierarchical trajectory. HiReT organizes each patient as an ordered sequence of visits, each visit as an ordered sequence of days, and each day as a collection of typed clinical evidence. Historical visits are complete, whereas the index visit is truncated at the landmark. This organization preserves disease evolution both within and across hospitalizations.

Evidence nodes. The i-th visit is represented as

$$V_i = \left(B_i, \{M_{i,j}, T_{i,j}, C_{i,j}, R_{i,j}\}_{j=1}^{D_i}\right), \tag{2}$$

where B_i is the visit background: patient attributes and encounter context shared across all days of that visit, and D_i is the number of 24-hour calendar days in visit i. The four dynamic types M, T, C, and R are recorded per day and denote measurements (physiological, laboratory, and microbiology observations), treatments (medications, procedures, and interventions), careflow (transfers and service trajectories), and reports (textual findings and impressions), respectively. The reconstruction therefore provides typed node sequences within each visit and an ordered visit sequence for subsequent relational and hierarchical modeling.

3.3 HIERARCHICAL RELATIONAL TRAJECTORY MODELING

Table 1: Concrete implementations of functions and operators in the model.

Symbol	Concrete implementation
Enc	Frozen Qwen3-Embedding-0.6B; L2-normalized 512-D last-token embedding.
ϕ_B	Separate $\mathrm{Linear}(512, 256)$–GELU–LayerNorm–Dropout(0.1) block for the visit background.
ϕ_k	Separate $\mathrm{Linear}(512, 256)$–GELU–LayerNorm–Dropout(0.1) block for each node type.
GRU_k	Four separate one-layer GRUs with 256-D input and hidden states, one for each of $M/T/C/R$.
$W_{s\to t}$	Separate $\mathrm{Linear}(256, 256)$ transform for each of the 16 directed relation types.
Q_t	Separate $\mathrm{Linear}(256, 256)$ query projection for each dynamic target type.
$f_{s\to t}$	Separate $\mathrm{Linear}(769, 1)$ scorer over message, target query, background, and source presence.
$\mathrm{GRUCell}_t$	Separate $\mathrm{GRUCell}(256, 256)$ update for each dynamic target type.
LN_t	Separate 256-D LayerNorm for each dynamic target type.
$\mathrm{maskedSoftmax}$	Parameter-free softmax over available sources or node types.
$W_k,\ f_D$	Five $\mathrm{Linear}(256, 256)$ type transforms and one shared $\mathrm{Linear}(513, 1)$ DayFuse scorer.
ψ	$\mathrm{Linear}(256, 256)$–GELU–LayerNorm–Dropout(0.1) day-update block.
$\mathrm{GRU}_D, \mathrm{GRU}_V$	One-layer $\mathrm{GRU}(256, 256)$ day and visit encoders with last-state readout.
$\mathrm{SmoothL1}$	Standard unit-transition quadratic-to-linear loss on the expected class position.

Table 1 specifies the implementation behind each operator used below.

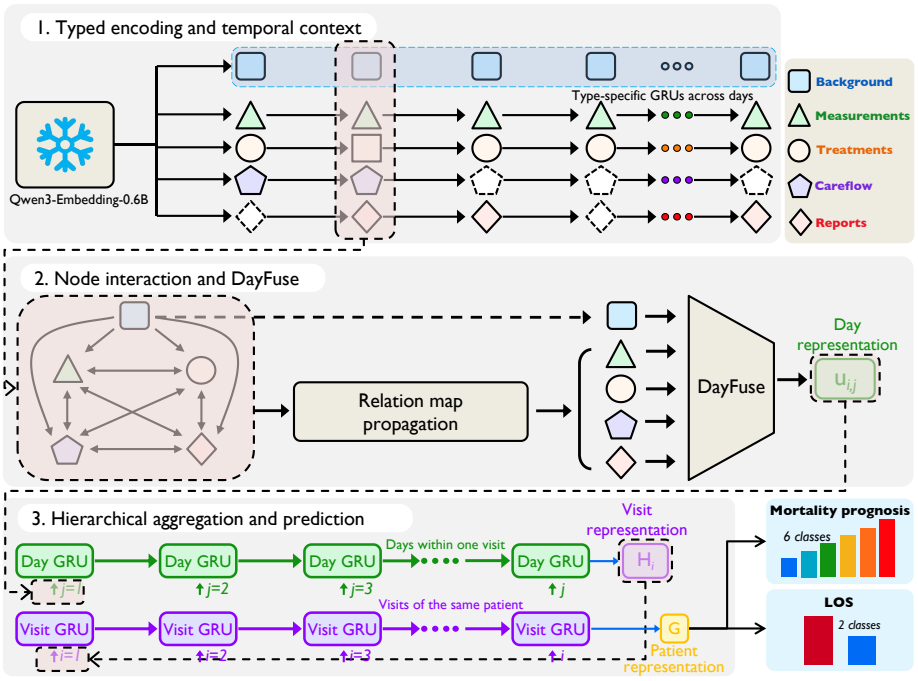

Figure 2: HiReT architecture. Frozen embeddings are projected into typed node streams and contextualized across days. Relation-map propagation carries visit background to the dynamic nodes and exchanges information among $M/T/C/R$. DayFuse produces day representations, followed by ordered day-to-visit and visit-to-patient aggregation and task-specific heads. Dashed shapes mark nodes unavailable on that day; Table 7 reports the availability of each node type.

Typed encoding and temporal context. Each textual node is encoded once by frozen Qwen3-Embedding-0.6B. Type-specific projectors map the normalized embeddings into a shared d-dimensional space:

$$\mathbf{z}_{B,i} = \phi_B(\mathrm{Enc}(B_i)), \qquad \mathbf{z}_{k,i,j} = \phi_k(\mathrm{Enc}(k_{i,j})), \quad k \in \mathcal{K} = \{M, T, C, R\}. \tag{3}$$

Here, i and j index the visit and day, k indexes a dynamic evidence type, and $\mathcal{K}$ is the set of four dynamic types. ϕ_B projects the visit background and ϕ_k projects the dynamic nodes of type k, as specified in Table 1; we set $d = 256$. This retains type-specific transformations while keeping the frozen encoder and its node features identical to the flat control. Availability masks remove absent nodes from all subsequent normalization and aggregation operations.

Let $r_{B,i,j} \in \{0, 1\}$ indicate whether the visit background is available on day j of visit i, and $r_{k,i,j} \in \{0, 1\}$ whether dynamic type k is available on that day. A type-specific GRU contextualizes each dynamic evidence stream along the valid days of a visit:

$$\mathbf{h}_{k,i,j} = \mathrm{GRU}_k(r_{k,i,j}\mathbf{z}_{k,i,j}, \mathbf{h}_{k,i,j-1}), \qquad \mathbf{z}^{(0)}_{k,i,j} = r_{k,i,j}\mathbf{h}_{k,i,j}. \tag{4}$$

Here, $\mathbf{h}_{k,i,j}$ is the internal recurrent state; on a day when type k is absent the input is zero and the exposed output is zeroed, so an absent node contributes no representation to that day; the recurrent state advances over every valid calendar day. Superscript (0) initializes relation-map propagation.

Node interaction and DayFuse. HiReT applies a typed relation map, which is fixed and directed

$$\mathcal{E} = \{B \to k : k \in \mathcal{K}\} \cup \{s \to t : s, t \in \mathcal{K},\ s \neq t\}. \tag{5}$$

Here, $\mathcal{E}$ is the set of directed relation types, s and t denote source and target types with $t \in \mathcal{K}$, and $\mathcal{N}(t)$ denotes the incoming source types of target t. Background informs every dynamic node and $M/T/C/R$ communicate pairwise. At propagation round ℓ, each relation transforms its source with relation-specific parameters and receives an input-dependent weight conditioned on the source

message, target state, and visit background; these parameters are shared across propagation rounds:

$$\begin{aligned}\mathbf{m}^{(\ell)}_{s\to t,i,j} &= W_{s\to t}\mathbf{z}^{(\ell-1)}_{s,i,j},\\ e^{(\ell)}_{s\to t,i,j} &= f_{s\to t}\Big([\mathbf{m}^{(\ell)}_{s\to t,i,j}\|Q_t\mathbf{z}^{(\ell-1)}_{t,i,j}\|\mathbf{z}_{B,i}\|r_{s,i,j}]\Big),\\ \alpha^{(\ell)}_{s\to t,i,j} &= \mathrm{maskedSoftmax}_{s\in\mathcal{N}(t)}\Big(e^{(\ell)}_{s\to t,i,j}; r_{s,i,j}\Big).\end{aligned} \tag{6}$$

Here, ℓ indexes the propagation round, $\mathbf{m}^{(\ell)}_{s\to t,i,j}$ is the message from s to t, $W_{s\to t}$ is its relation-specific transform, and $Q_t\mathbf{z}^{(\ell-1)}_{t,i,j}$ is the target query. The operator $\|$ denotes concatenation, $r_{s,i,j}\in\{0,1\}$ indicates whether source s is available, and $e^{(\ell)}_{s\to t,i,j}$ is the scalar output of the relation-specific scorer $f_{s\to t}$. The masked normalization gives $\alpha^{(\ell)}_{s\to t,i,j}$, the weight of source s among $\mathcal{N}(t)$; unavailable sources receive zero weight, and an all-masked source set returns a zero vector. The presence flag $r_{s,i,j}$ lets the shared scorer distinguish an available source from a zero-padded one. The same convention applies to DayFuse. For $s=B$, we use $\mathbf{z}^{(\ell-1)}_{B,i,j}\equiv\mathbf{z}_{B,i}$. A target-specific GRUCell integrates the weighted incoming message, followed by a residual connection and layer normalization:

$$\begin{aligned}\widetilde{\mathbf{m}}^{(\ell)}_{t,i,j} &= \sum_{s\in\mathcal{N}(t)}\alpha^{(\ell)}_{s\to t,i,j}\mathbf{m}^{(\ell)}_{s\to t,i,j},\\ \mathbf{z}^{(\ell)}_{t,i,j} &= r_{t,i,j}\,\mathrm{LN}_t\Big(\mathbf{z}^{(\ell-1)}_{t,i,j}+\mathrm{Dropout}\Big[\mathrm{GRUCell}_t(\widetilde{\mathbf{m}}^{(\ell)}_{t,i,j},\mathbf{z}^{(\ell-1)}_{t,i,j})\Big]\Big).\end{aligned} \tag{7}$$

where $\widetilde{\mathbf{m}}^{(\ell)}_{t,i,j}$ is the aggregated incoming message, and $\mathbf{z}^{(\ell)}_{t,i,j}$ is the updated target state passed to the next round. $\mathrm{GRUCell}_t$ controls how much of the incoming message is written to the state, LN_t is target-type-specific layer normalization, and $r_{t,i,j}$ is the availability mask of target t, which keeps an absent target at zero. L denotes the node-level depth of relation-map propagation applied at each day; we use two successive rounds before DayFuse. Finally, DayFuse combines the visit background with the relationally updated dynamic nodes into one day representation. With $\mathbf{q}_{B,i,j}=W_B\mathbf{z}_{B,i}$ and $\mathbf{q}_{k,i,j}=W_k\mathbf{z}^{(L)}_{k,i,j}$, it computes

$$\begin{aligned}\beta_{k,i,j} &= \mathrm{maskedSoftmax}_{k\in\{B,M,T,C,R\}}(f_D([\mathbf{q}_{k,i,j}\|\mathbf{z}_{B,i}\|r_{k,i,j}]); r_{k,i,j}),\\ \mathbf{u}_{i,j} &= \psi\left(\sum_{k\in\{B,M,T,C,R\}}\beta_{k,i,j}\mathbf{q}_{k,i,j}\right).\end{aligned} \tag{8}$$

where $\mathbf{q}_{k,i,j}$ is the DayFuse projection of type k, $\beta_{k,i,j}$ is its masked-normalized weight over the node types available on day j, ψ is the day-update network, and $\mathbf{u}_{i,j}$ is the resulting day representation passed to the day-level GRU. An absent type contributes no weight, so DayFuse never averages over missing evidence. The DayFuse scorer takes $2d+1=513$ inputs: the type projection, visit background, and scalar availability flag.

Hierarchical aggregation and prediction. A day GRU summarizes the ordered days of each visit, and a visit GRU then summarizes the ordered visits of the patient:

$$\begin{aligned}\mathbf{h}_{D,i,j} &= \mathrm{GRU}_D(\mathbf{u}_{i,j},\mathbf{h}_{D,i,j-1}), & \mathbf{H}_i &= \mathbf{h}_{D,i,D_i},\\ \mathbf{g}_i &= \mathrm{GRU}_V(\mathbf{H}_i,\mathbf{g}_{i-1}), & \mathbf{G} &= \mathbf{g}_N.\end{aligned} \tag{9}$$

$\mathbf{h}_{D,i,j}$ is the recurrent state after day j, $\mathbf{H}_i$ is the representation of visit i, $\mathbf{g}_i$ is the recurrent state after visit i, and $\mathbf{G}$ is the final patient representation. Both levels start from zero states and use the last valid state, which has accumulated the preceding sequence context; masks exclude padded days and visits. The resulting $\mathbf{G}$ jointly represents all available history up to the prediction time: the completed historical visits together with the observed portion of the index hospitalization.

Prediction and optimization. A linear head maps $\mathbf{G}$ to six mortality logits. Let $p_{n,c}$ be the softmax probability of class c, let $y_n\in\{0,\dots,C_{\mathrm{cls}}-1\}$ be the class label, where $C_{\mathrm{cls}}=6$, and let $\hat{y}_n=\sum_{c=0}^{C_{\mathrm{cls}}-1}c\,p_{n,c}$ be the expected class position and $\delta_n=\sum_{c=0}^{C_{\mathrm{cls}}-1}p_{n,c}|c-y_n|$ the expected distance to the true class. For a batch $\mathcal{B}$ of examples, we use

$$\mathcal{L}_{\mathrm{CE}}=-\frac{1}{|\mathcal{B}|}\sum_{n\in\mathcal{B}}\log p_{n,y_n},\quad \mathcal{L}_{\mathrm{ord}}=\frac{1}{|\mathcal{B}|}\sum_{n\in\mathcal{B}}\mathrm{SmoothL1}(\hat{y}_n-y_n),\quad \mathcal{L}_{\mathrm{dist}}=\frac{1}{|\mathcal{B}|}\sum_{n\in\mathcal{B}}\delta_n, \tag{10}$$

Table 2: Six-class mortality prediction; main results on the natural-distribution Test set (%). Learned models report mean ± standard deviation over three seeds; zero-shot models use a single run.

Method	Bal. Acc	Macro-F1	Macro-AUROC	QWK	Acc
RETAIN (Choi et al., 2016)	16.67±0.11	13.76±0.25	58.83±0.27	0.00±0.00	70.26±0.01
Med-BERT (Rasmy et al., 2021)	21.67±0.46	19.03±0.60	61.35±0.30	17.78±2.26	52.81±2.64
CEHR-BERT (Pang et al., 2021)	21.90±0.23	20.00±0.40	62.34±0.67	18.86±0.10	55.57±0.87
BEHRT (Li et al., 2020)	21.69±0.08	21.82±0.11	59.58±0.32	26.89±0.40	62.27±0.28
HiTANet (Luo et al., 2020)	18.49±0.21	15.27±0.17	58.06±1.48	8.73±1.20	59.96±2.26
Hi-BEHRT (Li et al., 2022)	21.38±0.12	20.54±0.24	60.93±0.32	25.04±0.18	59.25±0.52
Lingshu-I-8B (zero-shot) (Xu et al., 2025)	16.67	13.76	56.24	0.00	70.26
Qwen3.5-4B (zero-shot) (Qwen Team, 2026)	16.67	13.75	58.15	-0.02	70.25
Qwen3-4B (zero-shot) (Yang et al., 2025)	17.08	15.04	61.09	2.37	69.03
MedGemma-1.5-4B (zero-shot) (Sellergren et al., 2026)	17.25	2.59	50.93	0.03	4.04
HuatuoGPT-o1-8B (zero-shot) (Chen et al., 2024)	16.61	1.29	49.90	0.01	3.86
Qwen3-4B (SFT) (Yang et al., 2025)	22.51	22.04	77.87	29.75	70.61
Flat Embedding-0.6B	21.39±1.48	21.32±1.57	76.86±0.11	25.93±6.49	70.54±0.58
HiReT	**27.00**±0.35	**28.17**±0.41	**77.48**±0.11	**43.85**±0.89	**70.25**±0.14

$$\mathcal{L}_{\text{mort}} = \mathcal{L}_{\text{CE}} + 0.2\mathcal{L}_{\text{ord}} + 0.1\mathcal{L}_{\text{dist}}. \tag{11}$$

The ordinal term aligns the expected class position with the target, while the distance term penalizes probability assigned to more distant windows. Both terms are defined over the class index; we use $\beta_{\text{SmoothL1}} = 1.0$ in the SmoothL1 term and fixed weights 0.2 and 0.1 in all runs. For the LOS task, the same backbone feeds a two-class head that predicts whether total index-hospital LOS exceeds seven days, trained separately with cross-entropy over the two logits:

$$\mathcal{L}_{\text{LOS}} = -\frac{1}{|\mathcal{B}|}\sum_{n\in\mathcal{B}} \log p_{n,y_n}, \qquad y_n \in \{0,1\}, \tag{12}$$

where y_n indicates whether the LOS exceeds seven days and p_{n,y_n} is the softmax probability assigned to that label over the two classes. The LOS objective uses only two-class cross-entropy.

4 Experiments

4.1 Benchmarks

Cohort construction and splitting. We construct both naturally distributed benchmarks from MIMIC-IV (Johnson et al., 2023). Eligible patients are deterministically split 90/10 into patient-disjoint Train and Test sets before task-specific cohort selection. All eligible index-hospitalization examples are used for LOS, yielding 213,971 Train examples from 75,022 patients and 24,143 Test examples from 8,359 patients; positive rates (LOS $>$ 7 days) are 23.62% and 24.26% for Train and Test, respectively. From the same patient split, the mortality benchmark retains patients with an observed death event or sufficient follow-up to establish survival through 365 days, yielding 133,737 Train examples from 37,164 patients and 15,171 Test examples from 4,111 patients. Neither Test set is resampled; class counts, cohort details, and construction details appear in Appendix A.1.

For both tasks, index fields determined after the landmark, including discharge disposition and retrospective diagnosis summaries, are excluded. Raw images are unused; textual findings and impressions form R.

4.2 Baselines and Implementation

We compare representative methods from the two modeling lines reviewed in Section 2: linearized sequence modeling, which includes foundation models that operate on serialized records, and explicit hierarchy or relational modeling. Flat Embedding-0.6B is the strict architectural control: it uses the identical frozen Qwen3-Embedding-0.6B node features and examples as HiReT, without the hierarchy or relation map. All trainable baselines use the same examples and evidence, so methods differ in how they organize and encode this common input (Appendix A.2).

Hyperparameters and the full training and evaluation protocol, including seeds, checkpoint selection, and the inference setting for prompted models, are given in Appendix A.3.

4.3 Metrics

We use Balanced Accuracy as the primary metric and additionally report Macro-F1, Macro-AUROC, Quadratic Weighted Kappa (QWK) (Fleiss & Cohen, 1973), and Accuracy. Balanced Accuracy

Table 3: Six-class mortality prediction; architectural-component and evidence-node ablations under the common training protocol. Each variant is trained from initialization and reports mean ± standard deviation over three seeds.

Configuration	Bal. Acc	Macro-F1	Macro-AUROC	QWK	Acc
Architectural components					
Full	27.00±0.35	28.17±0.41	77.48±0.11	43.85±0.89	70.25±0.14
w/o Typed Node Encoding	24.77±0.32	25.74±0.29	76.06±0.44	37.57±0.82	67.79±0.83
w/o Node Interaction	26.42±0.18	27.23±0.55	76.50±0.73	43.32±0.64	67.40±1.28
w/o Day Hierarchy	26.38±1.08	27.12±0.49	76.47±0.27	42.13±2.26	67.22±0.95
w/o Visit Hierarchy	25.53±1.10	26.29±0.75	77.23±0.53	39.21±3.80	69.21±1.43
Evidence nodes					
w/o B	25.11±0.58	25.71±0.08	71.36±0.17	35.59±1.54	64.55±0.70
w/o M	26.15±0.43	27.55±0.29	77.10±0.45	40.59±2.02	68.38±0.44
w/o T	25.70±0.40	26.55±0.25	76.90±0.27	40.70±1.34	68.21±0.39
w/o C	26.05±0.54	27.78±0.54	77.17±0.45	42.21±1.87	68.26±0.30
w/o R	26.23±0.52	27.74±0.53	76.92±0.30	43.05±0.81	67.55±0.43

and Macro-F1 weight all classes equally, QWK penalizes errors spanning more distant outcome windows, and Accuracy measures correctness under the natural Test distribution.

4.4 Main Results

Table 2 shows that HiReT has the strongest class-balanced and ordinal performance. Relative to the strongest baseline, Qwen3-4B SFT, it improves Balanced Accuracy, Macro-F1, and QWK by 4.49, 6.13, and 14.10 points, respectively. Expressed as relative improvements over that baseline (the gain divided by the baseline score), these correspond to 19.95%, 27.81%, and 47.39%. Against the same-encoder flat control, the corresponding absolute gains are 5.61, 6.85, and 17.92 points, with higher recall across all five mortality windows (Appendix A.4). These results locate the main advantage in minority-class and ordinal prediction.

4.5 Hierarchy and Typed-Evidence Ablations

Table 3 isolates the model components. Removing Typed Node Encoding merges the available daily evidence before a shared projector, thereby eliminating type-specific encoding, interaction, and DayFuse. The remaining variants respectively bypass relation-map propagation, replace the day GRU with masked day averaging, or replace the visit GRU with masked visit averaging. Removing Typed Node Encoding lowers Balanced Accuracy, Macro-F1 and QWK by 2.23, 2.43, and 6.28 points, respectively; these are the largest losses among the architectural rows on these three metrics. This identifies clinically typed organization as the principal source of the model's balanced and ordinal gains. Removing the visit hierarchy causes the next-largest QWK loss (4.64 points), supporting the use of longitudinal encounter order, while removing the day hierarchy produces the largest Accuracy loss (3.03 points). Node Interaction yields smaller but consistent gains across all five metrics.

The evidence-node rows provide a clinically interpretable importance profile. Removing the visit background B causes the largest decline in every metric, showing that shared visit context is the dominant conditioning source. Among dynamic nodes, treatments T contribute most to Balanced Accuracy and Macro-F1, consistent with therapies and procedures reflecting clinicians' responses to patient state, while measurements M have the largest QWK effect among the dynamic nodes. Node-level availability on the Test set is reported in Table 7 (Appendix A.2).

4.6 Structural Use and Robustness

Table 4(A) applies inference-time interventions to the full model trained with seeds 42, 123, and 2026. Removing $B \to \{M, T, C, R\}$ degrades every metric more than disabling interaction within $\{M, T, C, R\}$. Together with the node ablations, this shows that background contributes both directly through DayFuse and indirectly by conditioning the dynamic streams. Disabling dynamic-node interaction also consistently degrades performance, supporting the value of cross-type exchange.

Removing or reversing historical visits causes the largest QWK reductions, while daily-node removal and cross-patient noise degrade all metrics; day-level interventions produce mostly small changes (Appendix A.8). Table 4(B) further shows that Flat Embedding is insensitive to visit re-

Table 4: Six-class mortality prediction; structural-use analyses under the common training protocol.

(A) Six-class mortality prediction; interventions on HiReT. (percentage-point change; mean ± standard deviation over three seeds). Negative values indicate degradation; no retraining is performed.

Intervention	ΔBal. Acc	ΔMacro-F1	ΔMacro-AUROC	ΔQWK	ΔAcc
Relation-path removal					
$B \rightarrow \{M, T, C, R\}$	-0.57±0.07	-0.98±0.41	-0.60±0.41	-1.71±0.34	-1.32±0.21
Interaction in $\{M, T, C, R\}$	-0.47±0.06	-0.64±0.11	-0.24±0.21	-0.60±0.72	-0.87±0.16
Input-structure perturbation					
Remove 50% history visits	-2.41±0.55	-2.43±0.59	-1.37±0.28	-7.19±1.51	-1.13±0.09
Reverse history visits	-1.65±0.03	-1.52±0.12	-1.08±0.02	-5.57±0.06	-1.79±0.11
Remove 50% nodes per day	-1.00±0.20	-1.45±0.15	-1.46±0.29	-2.26±1.41	-2.43±0.35
Inject 50% cross-patient noise	-1.58±0.30	-1.43±0.28	-1.22±0.04	-4.20±0.83	-2.14±0.17

(B) Six-class mortality prediction; cross-model perturbation audit. Each cell gives the absolute score with its clean-to-perturbed change in parentheses (percentage points), averaged over three seeds.

	HiReT		Flat Embedding-0.6B		Hi-BEHRT	
Perturbation	Bal. Acc	QWK	Bal. Acc	QWK	Bal. Acc	QWK
Remove 50% historical visits	24.59 (-2.41)	36.66 (-7.19)	21.05 (-0.34)	24.33 (-1.59)	20.97 (-0.41)	22.88 (-2.15)
Reverse historical visits	25.35 (-1.65)	38.28 (-5.57)	21.39 (+0.00)	25.93 (+0.00)	19.44 (-1.94)	16.96 (-8.08)
Remove 50% nodes per day	26.00 (-1.00)	41.58 (-2.26)	22.73 (+1.34)	31.61 (+5.68)	22.01 (+0.63)	28.18 (+3.14)
Inject 50% cross-patient noise	25.42 (-1.58)	39.65 (-4.20)	20.06 (-1.33)	19.26 (-6.66)	21.21 (-0.04)	25.87 (+1.03)

Table 5: Prolonged-LOS prediction; main results on the patient-disjoint, natural-distribution Test set (%). Learned multi-seed results report mean ± standard deviation; zero-shot models use a single run.

Method	Bal. Acc	Macro-F1	Macro-AUROC	QWK	Acc
RETAIN (Choi et al., 2016)	50.97±0.11	46.37±0.25	58.85±0.27	2.79±0.11	75.21±0.04
Med-BERT (Rasmy et al., 2021)	52.52±0.12	52.20±0.14	55.05±0.22	5.83±0.17	70.12±0.08
CEHR-BERT (Pang et al., 2021)	52.72±0.23	52.58±0.29	54.97±0.30	6.16±0.10	69.51±0.29
BEHRT (Li et al., 2020)	55.88±0.12	55.86±0.14	60.06±0.11	14.61±0.24	74.73±0.12
HiTANet (Luo et al., 2020)	51.70±0.31	47.58±0.33	66.59±0.41	4.88±0.19	75.69±0.34
Hi-BEHRT (Li et al., 2022)	55.46±0.08	55.58±0.10	58.16±0.09	12.97±0.19	73.07±0.08
Lingshu-I-8B (zero-shot) (Xu et al., 2025)	51.30	26.57	53.33	1.33	28.78
Qwen3-4B (zero-shot) (Yang et al., 2025)	51.14	51.11	52.24	2.24	63.47
Qwen3.5-4B (zero-shot) (Qwen Team, 2026)	50.05	19.66	59.15	0.05	24.35
MedGemma-1.5-4B (zero-shot) (Sellergren et al., 2026)	50.35	48.92	50.91	0.58	55.53
HuatuoGPT-o1-8B (zero-shot) (Chen et al., 2024)	49.99	43.09	49.50	0.02	75.73
Qwen3-4B SFT (Yang et al., 2025)	51.26	46.36	62.46	3.69	75.84
Flat Embedding-0.6B	53.46±0.74	50.95±0.87	73.11±1.09	9.68±1.16	76.29±0.14
HiReT	**59.08±0.84**	**59.91±1.18**	**75.94±1.68**	**22.97±1.60**	**77.64±0.23**

versal and that both comparison models may improve after node removal, indicating limited use of the intervened structure. HiReT remains the strongest method in absolute Balanced Accuracy and QWK under all perturbations that we report.

4.7 Prolonged-LOS Benchmark

Table 5 shows that HiReT achieves the highest value on all five metrics. Relative to the strict same-encoder flat control, it improves Balanced Accuracy by 5.62, Macro-F1 by 8.96, Macro-AUROC by 2.83, QWK by 13.29, and Accuracy by 1.35 points. The same hierarchical relational representation therefore transfers across both tasks. Fixed-backbone transfer results appear in Appendix A.7.

5 Conclusion

We presented HiReT, a hierarchical relational trajectory model that combines a `visit-day-node` representation with an intra-visit typed relation map. On naturally distributed MIMIC-IV outcomes, this relational hierarchy improves class-balanced and ordinal prediction over flat pooling with the same frozen node encoder, and the same representation supports both mortality-window classification and early LOS prediction. The results support the central premise of this work: aligning a prognosis model with the units in which clinical evidence is generated improves prediction and makes the model's reliance on evidence more directly checkable. Component ablations and pathway interventions show how visit order, day progression, typed evidence organization, and selected relational pathways contribute. Future work will study external validation, fairness, and safety under clinical distribution shift.

AI USE STATEMENT

Generative AI tools assisted writing and polishing, literature retrieval and discovery, research ideation and execution, and drafting sections of the paper. Specific uses included code implementation and revision, experiment monitoring, result checking and interpretation, consistency checks of equations and method descriptions, and figure creation and editing. The authors reviewed the AI-assisted content, inspected generated code, checked experimental claims against the underlying predictions and configurations, and verified cited sources. Final scientific judgments and decisions remained with the authors, who take responsibility for all text, claims, code, and figures in this work.

ETHICS STATEMENT

This study uses the credentialed, de-identified MIMIC-IV database and does not involve prospective intervention or direct patient contact. The proposed model is intended for research and clinical decision support rather than autonomous medical decision making. Any deployment would require independent validation, appropriate governance, and clinician oversight.

REPRODUCIBILITY STATEMENT

Reproduction details are provided in Appendix A.1 (cohort construction, labels, splits, and observation windows), Appendix A.2 (node definitions, missingness, and baseline input construction), and Appendix A.3 (model and training settings). Per-seed mortality results and patient-level confidence intervals are reported in Appendices A.5 and A.6; perturbation results appear in Appendix A.8. Release scope is described in Appendix A.11. Code and trained model checkpoints will be released upon publication.

REFERENCES

Ricardo MS Carvalho, Daniela Oliveira, and Catia Pesquita. Knowledge graph embeddings for icu readmission prediction. *BMC medical informatics and decision making*, 23(1):12, 2023.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

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. *Advances in neural information processing systems*, 29, 2016.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H Shah. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8(1):577, 2025.

Juan Du, Dajian Zeng, Zhao Li, Jingxuan Liu, Mingqi Lv, Ling Chen, Dan Zhang, and Shouling Ji. An interpretable outcome prediction model based on electronic health records and hierarchical attention. *International Journal of Intelligent Systems*, 37(6):3460–3479, 2022.

Shaza Elsharief, Saeed Shurrab, Baraa Al Jorf, L Julián Lechuga López, and Farah E Shamout. Medmod: Multimodal benchmark for medical prediction tasks with electronic health records and chest x-ray scans. *Proceedings of Machine Learning Research*, 287:1–23, 2025.

Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

Paul Grundmann, Jan Frick, Dennis Fast, Thomas Steffek, Felix Gers, Wolfgang Nejdl, and Alexander Löser. Clinibench: A clinical outcome prediction benchmark for generative and encoder-based language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5360–5378, 2026.

Pengcheng Jiang, Cao Danica Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *International Conference on Learning Representations*, volume 2024, pp. 16839–16866, 2024.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, et al. 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, 2024.

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, 2020.

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, 2022.

Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala, Jose G Moreno, and Lynda Tamine. Revisiting the mimic-iv benchmark: Experiments using language models for electronic health records. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pp. 189–196, 2024.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 647–656, 2020.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1903–1911, 2017.

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 825–832, 2020.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine learning for health*, pp. 239–260. PMLR, 2021.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026. Accessed: 2026-09-26.

Sushravya Raghunath, Alvaro E. Ulloa Cerna, Linyuan Jing, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26:886–891, 2020. doi: 10.1038/s41591-020-0870-z. URL https://www.nature.com/articles/s41591-020-0870-z.

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): 86, 2021.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Yuqi Si and Kirk Roberts. Three-level hierarchical transformer networks for long-sequence and multiple clinical documents classification. *arXiv preprint arXiv:2104.08444*, 2021.

Ikaro Silva, George Moody, Daniel J. Scott, Leo A. Celi, and Roger G. Mark. Predicting mortality of ICU patients: The PhysioNet/Computing in Cardiology Challenge 2012. In *Computing in Cardiology*, volume 39, pp. 245–248, 2012. URL https://moody-challenge.physionet.org/2012/.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3): AIoa2300138, 2024.

Linna Wang, Xinyu Guo, Haoyue Shi, Yuehang Ma, Han Bao, Lihua Jiang, Li Zhao, Ziliang Feng, Tao Zhu, and Li Lu. Crisp: A causal relationships-guided deep learning framework for advanced icu mortality prediction. *BMC medical informatics and decision making*, 25(1):165, 2025.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2839–2852, 2023.

Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health*, pp. 259–278. PMLR, 2022.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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, 2023.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489, 2016.

Wenfang Yao, Chen Liu, Kejing Yin, William K Cheung, and Jing Qin. Addressing asynchronicity in clinical multimodal fusion via individualized chest x-ray generation. *Advances in Neural Information Processing Systems*, 37:29001–29028, 2024.

A APPENDIX

CONTENTS

- A.1 Benchmark Construction
- A.2 Node and Baseline Construction
- A.3 Implementation and Evaluation Protocol
- A.4 Class-Wise Recall
- A.5 Per-Seed Main-Method Results
- A.6 Patient-Level Confidence Intervals
- A.7 Cross-Task Transfer to Prolonged LOS
- A.8 Complete Perturbation Results
- A.9 Encoder Control
- A.10 Subgroup Audit
- A.11 Scope and Reproducibility

A.1 BENCHMARK CONSTRUCTION

For each eligible index admission, completed prior admissions form historical visits and the index visit contains evidence before the 24-hour landmark.

Mortality-outcome ascertainment. Labels are assigned by a deterministic sequence of temporally disjoint rules. Class A denotes death after the landmark and no later than discharge from the index admission. For patients discharged alive, elapsed time from index discharge to death determines classes B–E: 0–30, 31–90, 91–180, and 181–365 days, respectively. Class F requires positive evidence of being event-free through day 365, established either by a recorded death after day 365 or by a subsequent observed hospitalization extending follow-up to or beyond that horizon. A missing death record alone is never treated as survival. Examples with follow-up ending before day 365, death at or before the landmark, or temporally inconsistent admission, discharge, and outcome records are excluded. The resulting six classes are mutually exclusive, and every retained example has an ascertainable 365-day outcome. All methods use this identical labeled cohort. All rolling examples from a patient remain in one partition.

Interpretation of the survival class. Mortality prediction distinguishes death from survival relative to a specified endpoint: the PhysioNet/CinC 2012 benchmark uses survival to hospital discharge (Silva et al., 2012), whereas one-year mortality prediction distinguishes patients who are dead versus alive at one year (Raghunath et al., 2020). Our class F follows this endpoint-specific interpretation, with the horizon set to 365 days after index discharge. A patient with a recorded death after this horizon and a patient observed alive through it have the same target status, even though their eventual outcomes may differ. We do not assign an unobserved death date to a patient who is alive at last follow-up. Death exactly 365 days after discharge belongs to class E; absence of a death record with follow-up covering at least 365 days supports class F. Consequently, F denotes *365-day survival*, not an additional observed death-time interval. Combining A–E yields the death category and F the survival category for this outcome horizon, while the six-class target additionally resolves the timing of death. This is classification with an ascertainable endpoint. Excluding insufficiently followed examples may select a cohort with greater observed healthcare contact; performance therefore pertains to the retained, outcome-ascertainable cohort.

The LOS task uses the same landmark. Admissions ending by 24 hours are excluded; deaths after the landmark remain and are labeled by observed LOS. The labels are total LOS ≤ 7 and > 7 days. Table 6 summarizes both benchmark cohorts, splits, and natural label distributions.

Index discharge fields, retrospective diagnosis blocks, outcome wording, and post-landmark medication end times are removed. Historical visits remain complete. Calendar-day aggregation and the resulting observation spans are described under Conservative landmark materialization below.

Eligibility and rolling landmarks. An admission is eligible as an index visit only when the patient has at least one completed earlier hospitalization and is alive and still hospitalized at the 24-hour landmark. Let K denote the total number of hospitalizations in a patient's recorded trajectory. Admissions 2 through K provide $K - 1$ candidate index visits; those satisfying the landmark and

Table 6: Cohort statistics for both tasks. Parenthesized mortality entries are class counts for A–F; LOS entries are counts for LOS ≤ 7 and > 7 days.

Task	Split	Patients	Examples	Label counts
Six-class mortality	Train	37,164	133,737	(4,660; 7,068; 8,869; 8,471; 11,783; 92,886)
	Test	4,111	15,171	(545; 765; 967; 948; 1,287; 10,659)
Prolonged LOS	Train	75,022	213,971	(163,425; 50,546)
	Test	8,359	24,143	(18,285; 5,858)

task-specific eligibility criteria become prediction examples. The permanent split is deterministic (seed 20260816), patient-level, and preserves natural prevalence; all examples from a patient remain in one partition.

The mortality Train and Test partitions have closely matched natural class proportions, with survival (F) accounting for 69.45% and 70.26% of examples, respectively. The prolonged-LOS partitions likewise preserve their natural majority class. Patient-level assignment prevents examples constructed from different admissions of the same patient from crossing partitions.

Conservative landmark materialization. The available source representation is aggregated by calendar day. We therefore retain an index-day aggregate only when its full calendar interval ends no later than the landmark. This strict rule prevents post-landmark evidence from entering the input, although the effective observation span can be shorter than 24 hours. Across examples it is 0–24 hours (mean 10.04) for mortality and 0.02–24 hours (mean 10.48) for LOS, and the effective observation span is stored with each example.

A.2 Node and Baseline Construction

B is the visit background: patient attributes and encounter context. M are measurements (physiological, laboratory, and microbiology observations); T treatments (medications, procedures, and interventions); C careflow (transfers and service trajectories); and R reports (textual findings and impressions). Raw image pixels are excluded. Each node is embedded once by Qwen3-Embedding-0.6B and cached. HiReT updates only its projectors, temporal encoders, relation map, hierarchical GRUs, and the prediction head.

Sparse evidence nodes and missingness. The evidence is aggregated into consecutive 24-hour calendar days from the visit start time, and D_i is the number of such days in visit i. For every valid calendar day, the input construction allocates logical slots for $M/T/C/R$, but a slot does not imply that the corresponding evidence was observed. Let $r_{k,i,j} \in \{0, 1\}$ indicate whether type $k \in \{M, T, C, R\}$ is recorded on day j of visit i. Empty fields and placeholder strings such as `None documented.` are treated as semantically missing: their feature slots remain zero and $r_{k,i,j} = 0$. Each dynamic type shares one type-specific GRU across the days of a visit; no GRU is instantiated for an individual missing node. Its masked temporal encoding is

$$\widetilde{\mathbf{z}}_{k,i,j} = r_{k,i,j}\mathbf{z}_{k,i,j}, \qquad \mathbf{h}_{k,i,j} = \mathrm{GRU}_k(\widetilde{\mathbf{z}}_{k,i,j}, \mathbf{h}_{k,i,j-1}), \qquad \mathbf{z}^{(0)}_{k,i,j} = r_{k,i,j}\mathbf{h}_{k,i,j}. \tag{13}$$

Here, $\mathbf{z}^{(0)}_{k,i,j}$ is the masked output used to initialize relation-map propagation, as in Eq. 4. Thus, on a day where type k is absent, the GRU receives a zero input and continues advancing its internal state along the valid calendar-day sequence, while the exposed state for that slot is reset to zero. A later observation of the same type is therefore encoded from the recurrent state carried through the preceding day sequence. In Node Interaction, absent sources receive zero relation weight and absent targets remain zero after updating. DayFuse normalizes only over the visit background and dynamic types observed on that day. The visit-level background B_i remains available throughout the visit.

Table 7 reports the node availability on the natural-distribution Test set. The visit background is present for every visit; measurements and treatments appear on every day, while careflow (22.1%) and reports (5.0%) are sparse. Days carrying all four dynamic types account for 2.9%, and two or three dynamic types are present on 75.8% and 21.3% of days, respectively.

Table 7: Node availability on the natural-distribution Test set (%). The visit background is counted per visit and the dynamic types per (visit, day) cell.

Node	Available (%)
Background (per visit)	100.00
Measurements (per day)	100.00
Treatments (per day)	100.00
Careflow (per day)	22.14
Reports (per day)	5.03

Every trainable mortality baseline, including Qwen3-4B SFT, uses the same 133,737 Train examples, and every method is evaluated on the same 15,171 Test examples. Patient identities, labels, visit histories, index landmarks, and raw $B/M/T/C/R$ evidence records are identical across methods. Structured baselines tokenize and organize this common evidence according to their native sequential or hierarchical representations. Flat Embedding-0.6B uses the identical cached node embeddings with global pooling, while prompted models receive a linearized rendering with constrained six-label instructions. Architecture-specific tokenization and context-length handling are applied only after constructing the common examples.

A.3 Implementation and Evaluation Protocol

Frozen Qwen3-Embedding-0.6B features are projected to hidden size 256. The learned network uses dropout 0.1, two relation-map rounds, one-layer day and visit GRUs, and the last valid recurrent state at both hierarchy levels. We optimize for 12 epochs with AdamW, batch size 16, learning rate 10^{-4}, weight decay 0.01, mixed precision, and gradient clipping at 1.0.

We evaluated several combinations of propagation depth and daily readout before fixing the configuration used in all reported experiments. Table 8 compares the resulting settings: two rounds with a last-state readout is the strongest combination on Balanced Accuracy, Macro-F1, Macro-AUROC and QWK, adding a third round does not improve the ordinal metrics (44.12 versus 44.86 QWK) and lowers accuracy, and replacing the last-state readout with attention or mean pooling costs 1.4–3.7 QWK points at the same depth. The reported model therefore uses two propagation rounds with a last-state readout.

Table 8: Six-class mortality prediction; combinations of propagation depth and daily readout (seed 42, %).

Setting	Bal. Acc	Macro-F1	Macro-AUROC	QWK	Acc
L=1, attention readout	26.69	27.35	76.46	42.02	69.86
L=2, attention readout	26.09	26.67	76.41	42.38	69.99
L=2, last-state readout (used)	27.12	27.92	77.60	44.86	70.29
L=2, mean readout	26.07	27.30	77.14	41.18	70.62
L=3, attention readout	26.73	27.14	77.07	44.12	69.01

Multi-seed results for trained mortality models are evaluated on seeds 42, 123, and 2026, and one-seed analyses are marked explicitly in the tables. Prompted zero-shot models use one deterministic inference run. They receive the linearized record with a fixed task-specific system prompt that lists the outcome labels and asks for exactly one letter (A–F for mortality and A–B for LOS). We do not sample text; we score the allowed label letters under the model's conditional distribution and take the highest-scoring letter. Inference uses bfloat16 and a maximum context length of 8,192 tokens, and all prompted predictions were parsed successfully (0% invalid). All reported results come from the final-epoch checkpoint of each learned model, evaluated on the Test set.

A.4 Class-Wise Recall

Table 9: Six-class mortality prediction; recall (%) for each outcome on the natural Test set. A–E are the five mortality windows and F denotes survival through 365 days. All trained models are averaged over three seeds; † denotes one run.

Method	A	B	C	D	E	F
HiReT	16.88	18.30	17.17	4.22	10.88	94.57
Flat Embedding-0.6B	3.43	13.59	7.69	0.21	5.57	97.87
RETAIN	0.00	0.00	0.00	0.00	0.00	100.00
Med-BERT	34.19	11.50	6.17	1.90	5.02	71.26
CEHR-BERT	28.87	10.85	7.34	4.43	4.69	75.22
BEHRT	4.53	4.36	6.96	9.11	21.06	84.10
HiTANet	23.30	3.79	0.00	0.00	0.00	83.88
Hi-BEHRT	1.77	2.27	6.83	11.46	26.75	79.20
Lingshu-I-8B (zero-shot)†	0.00	0.00	0.00	0.00	0.00	100.00
Qwen3.5-4B (zero-shot) (Qwen Team, 2026)†	0.00	0.00	0.00	0.00	0.00	99.99
Qwen3-4B (zero-shot)†	0.00	1.18	1.86	0.00	1.63	97.80
Qwen3-4B (SFT)	19.82	3.92	13.24	0.00	0.08	98.00

Table 9 shows that HiReT raises recall over the same-encoder flat control in every mortality window A–E while retaining 94.57% recall for survival. In contrast, RETAIN and two prompted models predict almost exclusively the majority survival class. This decomposition explains why similar overall Accuracy can coexist with large differences in Balanced Accuracy, Macro-F1, and QWK.

A.5 Per-Seed Main-Method Results

Table 10: Six-class mortality prediction; HiReT and the strict same-encoder flat control on the natural Test distribution (%). Mean and standard deviation are across three seeds.

Method	Seed	Acc	Bal. Acc	Macro-F1	Macro-AUROC	QWK
HiReT	42	70.20	27.12	27.92	77.60	44.86
	123	70.40	26.60	27.93	77.47	43.22
	2026	70.14	27.28	28.65	77.38	43.46
	Mean ± std	70.25±0.14	27.00±0.35	28.17±0.41	77.48±0.11	43.85±0.89
Flat Embedding-0.6B	42	69.88	23.10	23.12	76.98	33.41
	123	70.87	20.65	20.61	76.84	22.55
	2026	70.89	20.42	20.22	76.77	21.82
	Mean ± std	70.54±0.58	21.39±1.48	21.32±1.57	76.86±0.11	25.93±6.49

Table 10 shows that across the three seeds, HiReT is substantially more stable than the flat control on QWK (0.89 versus 6.49 standard deviation) and on the two class-balanced metrics. The per-seed results therefore support the aggregate main-result comparison without selecting a favorable run.

A.6 Patient-Level Confidence Intervals

Table 11: Six-class mortality prediction; patient-cluster paired bootstrap comparison of HiReT and the strict same-encoder Flat Embedding-0.6B control on six-class mortality. The reported effect is HiReT minus Flat in percentage points. Each interval uses 2,000 resamples of the 4,111 Test patients, retains all rolling examples for every sampled patient, and averages the seed-paired effect across seeds 42, 123, and 2026 within each resample.

Metric	Paired difference	Patient-bootstrap 95% CI
Balanced Accuracy	+5.61	[+4.73, +6.46]
Macro-F1	+6.85	[+5.74, +7.94]
Macro-AUROC	+0.62	[+0.05, +1.37]
QWK	+17.92	[+15.76, +20.03]
Accuracy	−0.29	[−0.97, +0.39]

A.7 Cross-Task Transfer to Prolonged LOS

We test whether the mortality-trained representation transfers to prolonged LOS without updating the hierarchical backbone.

Table 12: Prolonged-LOS prediction; seed-42 cross-task transfer (%). The latter two rows hold the complete mortality-task backbone fixed and train only an identical 514-parameter binary linear head on the frozen patient representation.

Configuration	Bal. Acc	Macro-F1	Macro-AUROC	QWK	Acc
LOS training from scratch	58.12	58.56	77.04	21.17	77.77
Mortality-pretrained backbone + linear head	54.17	52.45	69.69	11.39	76.18
Random backbone + linear head	51.12	45.79	67.09	3.30	75.99

Table 12 shows that, with an identical 514-parameter linear head, mortality pretraining improves over a random fixed backbone by 3.05 Balanced-Accuracy, 6.66 Macro-F1, 2.60 Macro-AUROC, 8.09 QWK, and 0.19 Accuracy points. End-to-end LOS training remains stronger by 3.95, 6.11, 7.35, 9.78, and 1.59 points on the same metrics, respectively; the frozen-backbone results nonetheless show that the mortality-learned representation transfers to the LOS task.

A.8 Complete Perturbation Results

All perturbations are applied to the patient-disjoint Test inputs without retraining. Reported changes are relative to the corresponding clean evaluation and are averaged over seeds 42, 123, and 2026.

Table 13: Six-class mortality prediction; complete HiReT test-time perturbation matrix (mean change in percentage points over three seeds).

Perturbation	ΔBal. Acc	ΔMacro-F1	ΔMacro-AUROC	ΔQWK	ΔAcc
Remove 50% historical visits	-2.41	-2.43	-1.37	-7.19	-1.13
Reverse historical visits	-1.65	-1.52	-1.08	-5.57	-1.79
Shuffle historical visits	-0.83	-0.71	-0.56	-2.87	-1.84
Remove 50% days	-0.53	-0.25	-0.48	-1.37	-2.05
Shuffle days	-0.03	-0.29	-0.67	-0.62	-2.82
Remove first day	-0.01	-0.08	-0.55	-0.77	-2.72
Remove last day	-0.08	-0.04	-0.02	-0.47	-2.02
Remove first and last days	-0.07	-0.17	-0.53	-0.91	-2.65
Remove 50% middle days	-0.27	-0.02	-0.04	-0.94	-1.71
Remove 50% nodes per day	-1.00	-1.45	-1.46	-2.26	-2.43
Shuffle nodes within each visit	-0.06	-0.28	-0.49	-0.58	-2.89
Remove all reports	-0.23	-0.18	-0.00	-0.71	-1.82
Inject 50% cross-patient noise	-1.58	-1.43	-1.22	-4.20	-2.14

Table 13 shows that the largest QWK changes occur after removing or reversing historical visits, while node removal and cross-patient noise affect all five metrics. Several localized day perturbations produce small changes in class-balanced metrics; the table is therefore interpreted as a map of which evidence manipulations the trained model uses most strongly.

Table 14 reports the mean incoming relation weight of the trained model, which gives a training-time view of the same question. The learned routing is far from uniform: measurements are the dominant source into treatments (64.7%) and treatments the dominant source into measurements (46.7%), while the visit background dominates the messages into careflow (56.7%). Reports contribute only 2–5% of the incoming weight of the other dynamic types, matching their 5.0% availability in the corpus. These weights are consistent with the ablations: B carries the largest share of the incoming weight and its removal causes the largest degradation in Table 3, whereas reports receive and send almost no weight and their removal is the mildest evidence-node change; the background-to-dynamic family is also the pathway whose removal degrades QWK most in Table 4(A).

Table 14: Six-class mortality prediction; mean learned incoming relation weight (%) on the natural Test set, averaged over seeds 42, 123 and 2026. Rows are destination types and columns are source types; the visit background receives no messages.

Destination	B	M	T	C	R
Round 1					
M	33.78	–	46.72	16.83	2.67
T	16.40	64.68	–	16.69	2.24
C	56.74	12.07	27.06	–	4.12
R	24.99	25.16	25.68	24.16	–
Round 2					
M	40.97	–	40.89	15.65	2.49
T	19.04	62.57	–	16.00	2.39
C	54.63	15.89	24.69	–	4.79
R	15.30	29.95	31.24	23.51	–

A.9 Encoder Control

Table 15: Six-class mortality prediction; frozen-encoder control (%).

Frozen encoder	Architecture	Bal. Acc	Macro-F1	Macro-AUROC	QWK	Acc
Qwen3-Embedding-0.6B	Flat	21.39	21.32	76.86	25.93	70.54
	HiReT	27.00	28.17	77.48	43.85	70.25
all-MiniLM-L6-v2	Flat	22.53	21.88	75.70	30.93	69.55
	HiReT	25.31	26.26	73.72	37.70	65.30

Table 15 shows that with Qwen3-Embedding-0.6B fixed, the hierarchical architecture provides large gains over flat pooling in Balanced Accuracy, Macro-F1, and QWK. The same direction holds on the class-balanced and ordinal metrics with the much smaller all-MiniLM encoder, showing that the benefit is not specific to one frozen representation model, although absolute performance varies with encoder quality.

Table 16 places the proposed configuration next to three alternatives that share the same frozen encoder, the same examples and the same hierarchy: a flat embedding, a day hierarchy without typed node modeling, and the same hierarchy with an attention-based day readout. The proposed configuration attains the highest Balanced Accuracy, Macro-F1, Macro-AUROC and QWK of the four rows.

Table 16: Six-class mortality prediction; alternatives under the same frozen Qwen3-Embedding-0.6B encoder (%). The day hierarchy without typed nodes is seed 42; the other rows report three-seed means.

Configuration	Bal. Acc	Macro-F1	Macro-AUROC	QWK	Acc
Flat embedding	21.39	21.32	76.86	25.93	70.54
Day hierarchy without typed nodes	24.77	25.74	76.06	37.57	68.79
Attention day readout	26.46	27.58	77.23	42.42	69.52
HiReT	27.12	27.92	77.60	44.86	70.29

A.10 SUBGROUP AUDIT

Table 17: Six-class mortality prediction; HiReT subgroup point estimates averaged over seeds 42, 123, and 2026. Samples counts index examples, so one patient may contribute several. Death sensitivity is the fraction of A–E examples predicted as non-survival, and survival specificity is the fraction of F examples predicted as F.

Dimension	Subgroup	Samples	Bal. Acc	Macro-AUROC	Death sens.	Survival spec.
Age	18–39	2,065	25.30	85.22	36.79	98.87
	40–59	4,961	27.67	78.19	48.94	92.21
	60–79	6,054	26.79	74.18	45.31	89.71
	≥ 80	2,091	24.93	68.28	43.00	84.14
Sex	Female	7,300	26.65	78.76	42.76	92.62
	Male	7,871	27.26	76.27	47.34	90.58
Race	Asian	398	29.63	75.49	63.94	85.61
	Black	2,921	25.66	81.53	41.72	96.14
	Hispanic/Latino	818	26.59	78.48	48.42	91.86
	Other/Unknown	704	28.03	75.71	47.55	89.16
	White	10,330	26.97	76.06	44.67	90.43
History	1–2 visits	6,002	26.01	79.89	36.90	95.47
	≥ 3 visits	9,169	27.48	76.21	50.35	88.96
Trajectory	Short (≤ 12 days)	5,384	24.61	80.91	31.38	96.58
	Medium (12–35 days)	4,850	28.09	78.43	47.35	91.40
	Long (> 35 days)	4,937	27.19	73.33	53.12	85.22

Table 17 shows that subgroup point estimates vary by demographic and trajectory characteristics. Sex has a relatively narrow Balanced-Accuracy range (26.65–27.26), whereas race spans 25.66–29.63 and age spans 24.93–27.67. Patients with at least three visits have higher death sensitivity than those with one or two visits (50.35 versus 36.90), and death sensitivity rises from 31.38 for short trajectories to 53.12 for long trajectories. These estimates are descriptive; smaller subgroups should be interpreted cautiously because of their limited sample sizes.

A.11 SCOPE AND REPRODUCIBILITY

We will release deterministic splits, manifests, field filters, cache construction, configurations, and evaluation scripts. MIMIC-IV remains under PhysioNet credentialing. These are retrospective single-database evaluations; external and prospective validation is required before clinical use.