HEAL-NET: HETEROGENEOUS EVIDENCE-ADAPTIVE LEARNING FOR MULTI-TASK CLINICAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic health records contain heterogeneous clinical evidence, including diagnoses, procedures, medications, laboratory measurements, and longitudinal encounter histories. Integrating these sources is challenging because they differ in semantic structure, clinical relevance, and reliability, while downstream prediction tasks require different uses of current and historical information. We introduce **HEAL-Net**, a **H**eterogeneous **E**vidence-**A**daptive **L**earning **Net**work for clinical prediction. HEAL-Net models patient records as visit-aware heterogeneous graphs with typed concepts, severity-aware labs, and knowledge-based relations. Its relational attention encoder and learnable temporal kernel integrate heterogeneous evidence across visits. Task-specific heads support mortality/readmission prediction, medication recommendation, and length-of-stay estimation through tailored historical fusion and output adjustments. Experiments on MIMIC-III/IV across four tasks show that HEAL-Net achieves the best reported mortality and readmission AUPRC/AUROC, improves AUPRC over GraphCare by 2.40–4.06 points, and attains the highest length-of-stay AUROC and medication sample-F1. Ablations validate graph structure, temporal encoding, and laboratory representation. Our code can be found in the Reproducibility statement.

1 INTRODUCTION

Electronic health records (EHRs) contain complementary evidence, including diagnoses, procedures, medications, laboratory measurements, and repeated encounters. These observations have distinct roles, and their relationships and temporal context can be as informative as the observations themselves (Choi et al., 2018; Ye et al., 2021; Chan et al., 2024). For example, a medication may gain meaning through its association with a diagnosis, while a recurring condition may have different implications across encounters. Effective clinical representation learning must therefore preserve observation roles, model their relationships, and adapt historical evidence to the prediction task (Choi et al., 2016a;b; Li et al., 2019). As illustrated in Figure 1, statistical analysis of the MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) datasets reveals a severe long-tailed disparity in the distribution of visit and task samples. Single-visit patients heavily dominate the database—accounting for roughly 60% to 80% of the population for tasks like length-of-stay (LOS) prediction and readmission/mortality assessment—which creates extreme data sparsity for longitudinal trajectories.

Existing EHR models use recurrent networks, Transformers, or graphs to learn longitudinal and concept-level representations (Choi et al., 2016a; Li et al., 2019; Choi et al., 2018; Chan et al., 2024). However, they often overlook typed relationships across encounters and treat biomedical knowledge as universally applicable, despite its context dependence and varying evidence availability (Ye et al., 2021; Jiang et al., 2024; Lu et al., 2022). In addition, generic prediction heads may not adequately capture the distinct structures of binary, multi-label, and ordinal clinical outcomes (Chan et al., 2024; Harutyunyan et al., 2017). These limitations motivate a unified, relation-aware framework with context-conditioned knowledge integration and task-specific prediction heads.

We propose **HEAL-Net**, a **H**eterogeneous **E**vidence-**A**daptive **L**earning **Net**work. HEAL-Net constructs patient-specific graphs containing encounter-level diagnoses, procedures, medications, and abnormal laboratory findings. Semantic, type, and severity embeddings encode clinical information, while clinical, temporal, and knowledge-supported relations connect concepts. A relation-aware graph encoder uses type-specific transformations and shared normalization across incoming edges,

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.

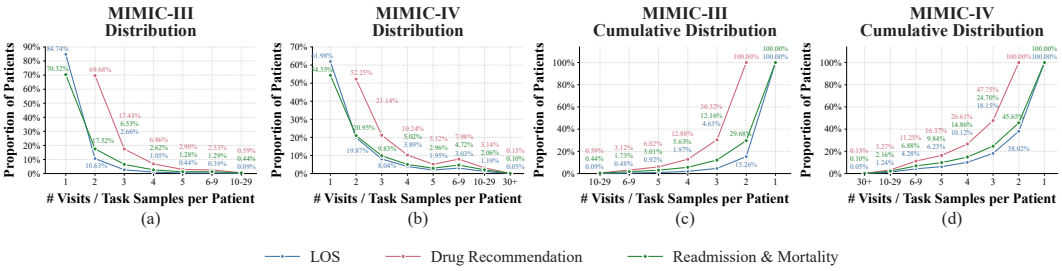

Figure 1: Distribution of # Visit / Task Samples in MIMIC-III and MIMIC-IV.

allowing heterogeneous evidence to compete contextually. Node representations are pooled into visits and contextualized through time-aware attention, with a shared exponential kernel modeling recency. Task-specific heads incorporate history for mortality, readmission, medication recommendation, and LOS, while a history-availability mask removes historical corrections when no prior encounter exists.

We evaluate HEAL-Net on mortality, readmission, medication recommendation, and LOS prediction using MIMIC-III and MIMIC-IV. It achieves the highest reported AUPRC and AUROC for both binary tasks on both datasets, improving AUPRC over GraphCare by 2.40–4.06 percentage points. It also obtains the highest LOS AUROC and medication sample-F1 on both datasets, although it does not lead every metric. Ablation studies support the benefits of heterogeneous graph structure, cross-visit encoding, and severity-aware laboratory features.

Our contributions are:

- **Visit-aware heterogeneous representation:** encounter-level graphs integrate typed concepts, laboratory findings, and knowledge-supported relations while preserving local context.
- **Evidence-aware relational attention:** shared incoming-edge normalization enables clinical, temporal, and knowledge-supported messages to compete adaptively.
- **Coordinated temporal aggregation:** a shared learnable kernel links graph-level, visit-level, and historical temporal modeling.
- **Task-adaptive historical integration:** availability-aware heads support binary, multi-label, and ordinal outcomes through bounded corrections, gated fusion, and order-preserving score adjustments.

2 Related Work

Representation Learning from Electronic Health Records. Early deep-learning approaches to EHR prediction used recurrent architectures to model longitudinal patient histories. Doctor AI demonstrated recurrent prediction of future clinical events from encounter sequences (Choi et al., 2016a). RETAIN introduced reverse-time attention to identify influential visits and variables in clinical prediction (Choi et al., 2016b). Transformer-based approaches, including BEHRT and Med-BERT, further developed contextual representations of structured medical records (Li et al., 2019; Rasmy et al., 2021). ***Analysis:*** These approaches establish the value of longitudinal representation learning. HEAL-Net complements this perspective by explicitly representing typed clinical relationships within and across encounters before aggregating visit representations.

Graph-Based Clinical Modeling. Graph-based EHR models exploit dependencies among medical concepts rather than treating them as independent features. GRAM incorporates medical ontologies through graph-based attention (Choi et al., 2017), while G-BERT combines graph-informed medical representations with Transformer-based encounter modeling for medication recommendation (Shang et al., 2019a). More generally, graph attention networks provide adaptive neighborhood aggregation (Veličković et al., 2018), and heterogeneous graph Transformers introduce type- and relation-dependent transformations (Hu et al., 2020). ***Analysis:*** HEAL-Net uses relation-specific projections while normalizing attention jointly across incoming relation types. Knowledge evidence enters as an attribute of graph relations, enabling its contribution to depend on both clinical context and node content.

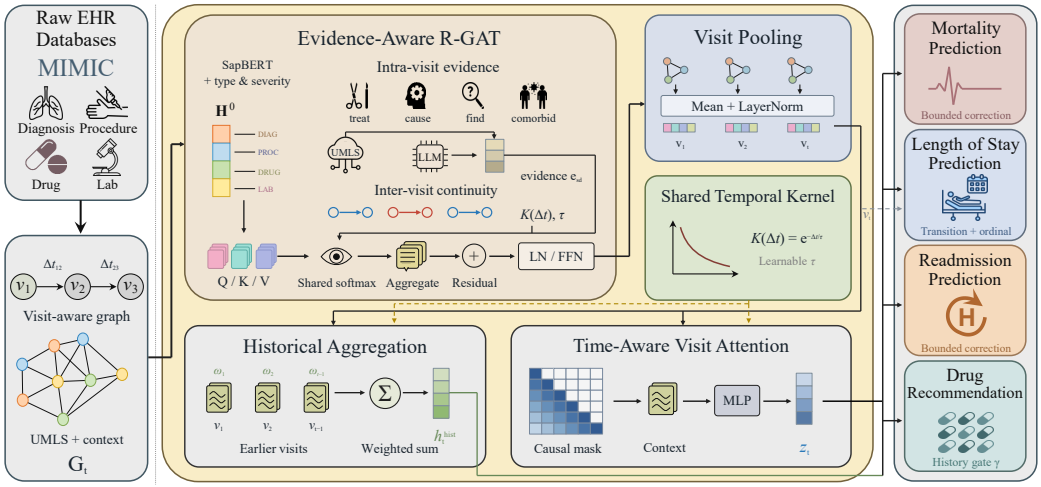

Figure 2: HEAL-Net architecture.

Biomedical Knowledge and Contextual Evidence. Biomedical knowledge can support representation learning by linking concepts that are sparsely observed or inconsistently coded. Concept encoders such as SapBERT align biomedical synonyms and provide reusable representations of medical terminology (Liu et al., 2021). ***Analysis:*** Knowledge applicability nevertheless depends on context. A general association between a condition and a medication does not imply that the association is relevant to every patient or encounter. HEAL-Net distinguishes structured path support from context-conditioned evidence and records whether the relevant contextual information is available.

Clinical Prediction with Different Output Structures. Clinical prediction includes binary outcomes, multi-label treatment sets, and ordered outcome categories. A shared encoder can support these tasks, but their output structures impose different requirements. Binary outcomes require calibrated risk estimates; medication recommendation involves correlated treatment labels; length of stay has an ordered structure that should be preserved in prediction. ***Analysis:*** HEAL-Net combines a shared representation architecture with prediction heads designed for these distinct label structures. Its ordinal head follows a cumulative-probability formulation with ordered thresholds, consistent with the rank-consistent modeling principles studied by Cao et al. (2020).

3 Methodology

HEAL-Net combines clinical observations, biomedical evidence, and longitudinal patient history. Figure 2 illustrate the network architecture for HEAL-Net. It first represents eligible observations as a visit-aware heterogeneous graph and aggregates clinical information through evidence-aware attention. The resulting node representations are pooled by visit and contextualized using time-aware attention. Finally, task-specific heads adapt how historical information contributes to each prediction.

3.1 Problem Formulation

Consider a patient with encounter history $\mathcal{H} = \{(\mathcal{X}_j, d_j)\}_{j=1}^{t}$, where $\mathcal{X}_j$ contains the clinical observations associated with encounter j, d_j is its reference timestamp, and t is the prediction encounter. For each task k, let $\mathcal{H}^{(k)}$ denote the history restricted to information available by its prediction cutoff. This restriction applies to clinical observations, source text, and the information used to construct graph edges and knowledge evidence.

Given $\mathcal{H}^{(k)}$, our goal is to predict the target $y^{(k)}$ for the current encounter. We consider four tasks: mortality and readmission with binary targets in $\{0, 1\}$; drug recommendation with a multi-label target in $\{0, 1\}^{|\mathcal{D}|}$; and length-of-stay prediction with an ordinal target in $\{1, \ldots, C\}$. Here, $\mathcal{D}$ is the drug vocabulary and C is the number of ordered length-of-stay categories.

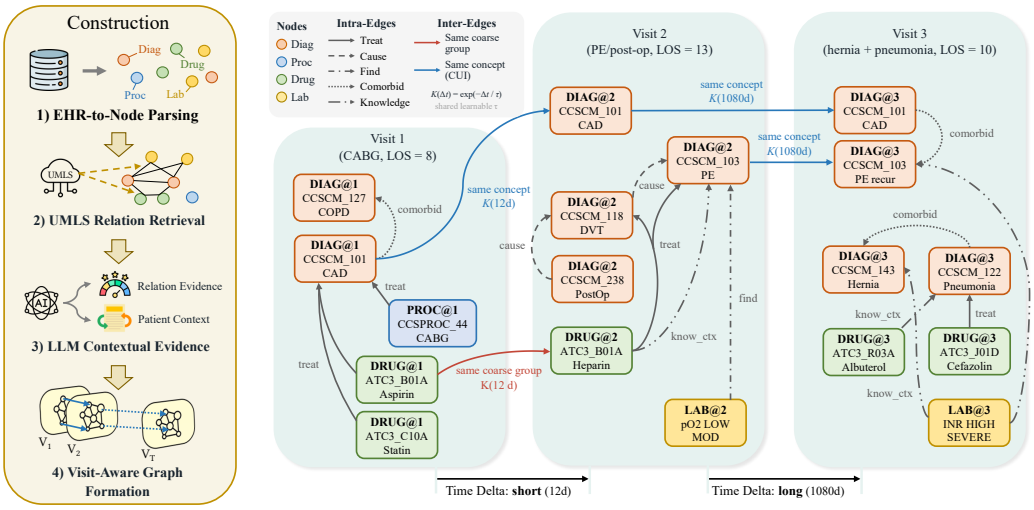

Figure 3: Heterogeneous clinical graph construction and example.

HEAL-Net adopts a history encoder f_θ with the same architecture across tasks and task-specific prediction functions g_ϕ:

$$\hat{\mathbf{y}}^{(k)} = g_{\phi_k}\left(f_{\theta_k}(\mathcal{H}^{(k)})\right).$$

The encoder and prediction head are trained independently for each task, using only inputs eligible under its task-specific cutoff. The output is a binary probability, medication-wise probabilities, or an ordinal class distribution, depending on the task.

3.2 Heterogeneous Clinical Graph

As shown in Figure 3, for each prediction instance, we construct a heterogeneous clinical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{R})$ with node types $\mathcal{T} = \{\texttt{DIAG}, \texttt{PROC}, \texttt{DRUG}, \texttt{LAB}\}$ and relation types $\mathcal{R}$. Each node is encounter-specific: the same concept in different encounters is represented by distinct nodes, preserving local context while allowing cross-visit connections.

HEAL-Net defines node identities with coarse-grained clinical codes to unify related concepts and reduce coding heterogeneity. Diagnoses and procedures are mapped from `ICD-9/10` to `CCS`, and medications from `NDC` to `ATC-3`. Laboratory measurements are discretized into analyte-specific severity categories $s_v \in \{\texttt{NORMAL}, \texttt{MILD}, \texttt{MODERATE}, \texttt{SEVERE}, \texttt{CRITICAL}\}$, with `NORMAL` results omitted to emphasize abnormal findings rather than measurement volume. Clinical concepts are additionally aligned to UMLS `CUIs` to obtain SapBERT representations and biomedical relations. Node features concatenate semantic, type, and severity embeddings:

$$\mathbf{x}_v = [\text{SapBERT}(c_v) \| \text{TypeEmb}(t_v) \| \text{SevEmb}(s_v)],$$

where c_v is the clinical concept, $t_v \in \mathcal{T}$ is the node type, and non-laboratory nodes use $s_v = \texttt{NA}$. These features are linearly projected to obtain the initial representations $\mathbf{h}_v^{(0)}$.

Edges $\mathcal{E} = \mathcal{E}_{\text{intra}} \cup \mathcal{E}_{\text{inter}}$ capture complementary intra-visit biomedical knowledge and inter-visit longitudinal continuity. Intra-visit edges comprise UMLS-augmented links retrieved through predefined relation types or paths, including associations not explicitly recorded in the EHR, and LLM-induced links derived from cutoff-eligible patient context to capture additional relevant relationships.

Elapsed time is modeled by a learnable exponential kernel:

$$K(\Delta t) = \exp(-\Delta t/\tau), \qquad \tau = \text{softplus}(\rho) + \varepsilon.$$

where $\Delta t \geq 0$, the trainable parameter ρ controls the decay timescale, and $\varepsilon > 0$ ensures a positive denominator. All timestamps use the same unit. The scalar τ is shared across graph temporal features, visit attention, and historical aggregation, providing a unified recency model throughout HEAL-Net.

A temporal edge between nodes j and i from different visits is created when the two nodes share either the same coarse-grained clinical concept group or the same UMLS `CUI`. Specifically, let $I_{\text{coarse}} = \mathbf{1}[\text{coarse}(j) = \text{coarse}(i)]$ and $I_{\text{CUI}} = \mathbf{1}[\text{CUI}_j = \text{CUI}_i]$, where $\text{coarse}(\cdot)$ maps a node to its coarse-grained clinical concept group. Both entries are zero for non-temporal relations. For each temporal edge, its weight is $w_{ji}^{\text{temp}} = K(\Delta t_{ji})$, and its temporal evidence vector is

$$\mathbf{e}_{ji}^{\text{temp}} = [K(\Delta t_{ji}), I_{\text{coarse}}, I_{\text{CUI}}] .$$

This formulation separates two concepts that were previously entangled: semantic continuity is described by the two indicators, while temporal persistence is described by the shared kernel. The kernel is computed in log space as $\log K(\Delta t_{ji}) = -\Delta t_{ji}/\tau$ to avoid numerical underflow. The fixed-width edge vector combines applicable knowledge and temporal features, with unavailable blocks zero-filled and accompanied by availability indicators.

The knowledge extraction procedure, relation schema, and continuity rule are detailed in Appendix C. They are established using only training data and external resources, without access to test labels.

3.3 Relational Graph Attention

The relational graph attention (R-GAT) learns which neighboring observations are most useful for representing each clinical concept. For an edge from source s to destination d with relation r, a learned score combines the source and destination representations with the edge feature. At each layer, attention score and aggregation are

$$a_{sd}^{r} = \text{LeakyReLU}\big(\mathbf{a}_r^{\top}[W_{Q,r}\mathbf{h}_d \| W_{K,r}\mathbf{h}_s \| \mathbf{e}_{sd}]\big) ,$$

$$\alpha_{sd}^{r} = \frac{\exp(a_{sd}^{r})}{\sum_{(s',r')\in\mathcal{N}(d)} \exp(a_{s'd}^{r'})}, \qquad \mathbf{m}_d = \sum_{(s,r)\in\mathcal{N}(d)} \alpha_{sd}^{r} W_{V,r}\mathbf{h}_s .$$

Here, $\mathbf{a}_r$ is a scoring vector and $\mathcal{N}(d)$ contains all incoming edge instances. Relation-specific projections allow clinical relationships to be interpreted differently, while normalization over all incoming edges lets these relationships compete in a common attention distribution. Thus, an edge's contribution depends on clinical content and supporting evidence, rather than its relation type alone.

The R-GAT uses multi-head attention, concatenates and projects the head outputs, and applies residual connections, dropout, layer normalization, and a feed-forward block. Isolated nodes receive zero aggregated messages but retain their information through the residual pathway. After the final graph layer, node representations are mean-pooled within each encounter and layer-normalized to obtain a visit embedding $\mathbf{v}_t$. This produces a fixed-dimensional representation without an additional learned node-selection mechanism. Empty visits use a learned vector.

3.4 Temporal Visit Attention

Visit embeddings are contextualized through causal multi-head attention. For one attention head, the score from visit j to visit i, weights, and outputs are computed as

$$s_{ji} = \frac{\langle W_Q\mathbf{v}_i, W_K\mathbf{v}_j\rangle}{\sqrt{d_k}} - \frac{\Delta d_{ji}}{\tau}, \quad j \le i; \quad \beta_{ji} = \frac{\exp(s_{ji})}{\sum_{j'\le i}\exp(s_{j'i})}, \quad \mathbf{o}_i = \sum_{j\le i}\beta_{ji} W_V\mathbf{v}_j .$$

Here, d_k denotes the key dimension, and τ is the same parameter as that used in the graph temporal kernel. Each visit attends only to itself and earlier visits, while the attention score combines a content-based compatibility term with a recency penalty governed by the decay timescale. Scores are softmax-normalized over eligible visits and used to aggregate their projected values; future visits are masked. Residual attention and feed-forward blocks produce contextualized visit embeddings. An MLP maps the prediction encounter's contextualized embedding to the patient representation $\mathbf{z}_t$.

The prediction heads also receive an explicit historical summary of earlier visits

$$\omega_{jt} = \frac{K(\Delta d_{jt})}{\sum_{i<t} K(\Delta d_{it})}, \qquad \mathbf{h}_t^{\text{hist}} = \sum_{j<t}\omega_{jt}\mathbf{v}_j .$$

This summary uses pre-attention visit embeddings and assigns greater weight to recent encounters. For patients without previous visits, it is set to zero. A history-availability indicator m_t is one when earlier visits exist and zero otherwise, allowing each head to disable its explicit historical adjustment when no history is available.

3.5 TASK-ADAPTIVE PREDICTION

Mortality and readmission. Let $k \in \{\text{mort}, \text{readm}\}$ index the mortality and readmission tasks, respectively. For each binary outcome, history provides a bounded correction to the base logit:

$$\ell^k_{t,\text{base}} = \mathbf{w}_k^\top \mathbf{z}_t + b_k, \quad \Delta^k_t = m_t B_k \tanh\big(f_k([\mathbf{z}_t \| \mathbf{h}^{\text{hist}}_t])\big), \quad \hat{y}^k_t = \sigma\big(\ell^k_{t,\text{base}} + \Delta^k_t\big).$$

Here, $B_k > 0$ is selected on the validation set, so $|\Delta^k_t| \le B_k$. History also affects the base prediction through $\mathbf{z}_t$, which is computed by the R-GAT and incorporates earlier encounters.

Drug recommendation. Since historical treatments could matter only for some patients, we project the historical summary into the patient representation space and fuse it via an elementwise gate:

$$\mathbf{r}_t = W_H \mathbf{h}^{\text{hist}}_t + \mathbf{b}_H, \quad \boldsymbol{\gamma}_t = m_t \sigma(W_\gamma[\mathbf{z}_t \| \mathbf{r}_t] + \mathbf{b}_\gamma), \quad \mathbf{z}^{\text{drug}}_t = (\mathbf{1} - \boldsymbol{\gamma}_t) \odot \mathbf{z}_t + \boldsymbol{\gamma}_t \odot \mathbf{r}_t.$$

A linear layer with independent sigmoids then predicts medication probabilities. When history is unavailable, $m_t = 0$ and the head uses $\mathbf{z}_t$ directly. Target medications are excluded from the input; historical medications are retained only when allowed by the task-specific cutoff.

Length of stay. To capture changes from historical evidence, we define $\mathbf{d}_t = \mathbf{v}_t - \mathbf{h}^{\text{hist}}_t$ and predict an ordinal score:

$$\eta_t = \mathbf{w}_{\text{los}}^\top \mathbf{z}_t + b_{\text{los}} + m_t B_{\text{los}} \tanh\big(f_{\text{los}}([\mathbf{z}_t \| \mathbf{h}^{\text{hist}}_t \| \mathbf{d}_t])\big), \quad q_{t,c} = P(y^{\text{los}}_t > c) = \sigma(\eta_t - b_c).$$

The thresholds satisfy $b_1 = \zeta_1$ and $b_c = b_{c-1} + \text{softplus}(\zeta_c)$ for $c > 1$, ensuring ordered cumulative probabilities. Class probabilities are obtained from adjacent differences, with $q_{t,0} = 1$ and $q_{t,C} = 0$. Since history modifies one shared ordinal score rather than individual thresholds, ordinal consistency is preserved.

Training. Mortality and readmission use binary cross-entropy. Drug recommendation uses BCE averaged over drugs, whereas LOS uses BCE over the $C - 1$ cumulative targets $\mathbb{I}[y^{\text{los}}_t > c]$. The joint objective is a weighted sum of task losses, each averaged over examples with available labels; unlabeled tasks contribute zero. Examples are constructed with task-specific cutoffs. Shared parameters are updated across task batches, while each head is updated only on its eligible inputs and labels.

4 EXPERIMENTS

Please see detailed experimental setup in Appendix B.

4.1 MAIN RESULTS

The main results are reported in Table 1 (Mortality and Readmission), Table 2 (Length of Stay), and Table 3 (Drug Recommendation). Across all four tasks and on both MIMIC-III and MIMIC-IV, HEAL-Net with the full TS+Lab+Graph configuration consistently delivers the strongest overall performance, validating our central thesis that a temporally unrolled, heterogeneous, relation-aware clinical graph provides a unified inductive bias that generalizes across acute, ordinal, and combinatorial clinical objectives.

Acute binary tasks (Mortality, Readmission). HEAL-Net achieves the best AUPRC and AUROC on both tasks across MIMIC-III and MIMIC-IV. The improvements over GraphCare—the strongest knowledge-graph baseline—are $+4.06$/$+3.39$ and $+2.40$/$+3.00$ (AUPRC/AUROC) on Mortality, and $+3.40$/$+2.21$ and $+3.48$/$+2.83$ on Readmission. Mortality is dominated by acute physiology: the TS+Lab variant already approaches the full model, and the graph adds a smaller but consistent margin via the bounded `dir_I` correction that prevents long-horizon context from overriding the present state. Readmission, in contrast, requires balancing therapeutic trajectory and pre-discharge stability, and is the regime where the relational graph contributes the largest absolute lift over sequence-only baselines.

Length of Stay. On the 10-class ordinal task, HEAL-Net leads on AUROC, Cohen's κ, and Accuracy on both datasets, with gains of $+1.71$ AUROC and $+2.81$ κ on MIMIC-III and $+1.64$ AUROC and $+2.48$ κ on MIMIC-IV over the strongest baseline (Transformer+Lab). Macro-F1 is matched within 0.5 on MIMIC-III and trails CNN+Lab by 0.34 on MIMIC-IV; we attribute the ordering to the

Table 1: Experimental Results for Mortality (Task 1) and Readmission (Task 2) Prediction.

Model	Input			Task 1: Mortality Prediction				Task 2: Readmission Prediction			
	TS	Lab	Graph	MIMIC-III		MIMIC-IV		MIMIC-III		MIMIC-IV	
				AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
CNN (LeCun et al., 1998)				10.87	57.44	2.87	62.26	68.06	62.29	69.00	68.19
CNN (LeCun et al., 1998)		√		10.53	57.56	/	/	73.00	68.97	70.35	68.00
RNN (Elman, 1990)				12.26	62.48	2.97	64.28	69.72	65.22	70.29	69.51
TCN (Bai et al., 2018)				8.28	54.35	2.91	62.50	70.94	66.47	67.73	67.52
Transformer (Vaswani et al., 2017)				12.31	59.17	2.51	61.86	68.77	63.78	68.11	67.42
Transformer (Vaswani et al., 2017)		√		10.39	61.05	2.57	64.31	62.66	68.96	70.13	67.58
Deepr (Nguyen et al., 2017)				10.68	61.76	2.68	63.35	68.73	65.18	68.91	68.02
Agent (Gao et al., 2020a)				9.62	56.75	2.53	62.03	70.04	66.90	69.14	68.91
AdaCare (Ma et al., 2020a)				10.06	59.63	3.14	63.69	71.95	67.02	68.84	68.55
ConCare (Ma et al., 2020b)				10.10	58.72	4.01	66.44	68.08	63.72	69.88	69.41
GRASP (Zhang et al., 2021)				9.90	60.44	3.61	66.92	70.20	66.07	70.02	69.36
RETAIN (Choi et al., 2016b)				10.07	52.32	2.78	62.91	66.32	64.35	69.20	68.95
StageNet (Gao et al., 2020b)				12.24	58.97	3.46	66.19	66.67	64.30	69.15	68.74
DeepSeek-zero shot (DeepSeek-AI, 2025)				7.92	47.09	1.86	50.03	51.91	49.46	51.84	50.81
DeepSeek-few shot (DeepSeek-AI, 2025)				7.87	48.39	1.91	50.87	52.20	50.04	52.00	51.12
MedTok (Su et al., 2025)	√			9.22	58.84	1.26	54.97	65.91	62.24	39.47	58.83
TRANS (Chen et al., 2024)	√		√	11.36	63.14	/	/	62.98	60.59	/	/
MultEHR (Chan et al., 2024)			√	15.64	72.71	2.68	64.16	63.95	60.96	66.30	66.77
GraphCare (Jiang et al., 2024)	√		√	15.80	70.30	4.70	71.50	70.50	68.90	70.10	69.80
HEAL-Net				14.11	63.10	3.13	69.04	67.87	66.75	71.28	70.37
		√		16.02	68.09	4.31	71.73	73.63	70.18	71.51	71.14
			√	12.67	64.90	3.31	69.86	69.71	68.11	71.93	70.99
		√	√	15.75	70.11	4.75	72.10	71.39	68.37	72.60	71.35
	√			14.73	69.52	5.16	70.43	67.76	67.00	71.29	70.80
	√	√		17.82	71.03	6.83	73.77	73.68	70.23	72.37	71.90
	√	√	√	**19.86**	**73.69**	**7.10**	**74.50**	**73.90**	**71.11**	**73.58**	**72.63**

Table 2: Experimental Results for Length of Stay Prediction (Task 3).

Model	Input			MIMIC-III				MIMIC-IV			
	TS	Lab	Graph	AUROC	Kappa	Accuracy	F1-score	AUROC	Kappa	Accuracy	F1-score
CNN (LeCun et al., 1998)				82.94	32.33	43.56	32.67	82.81	31.11	39.35	35.16
CNN (LeCun et al., 1998)		√		84.50	35.10	46.30	35.19	85.73	35.91	43.45	39.76
RNN (Elman, 1990)				83.76	32.43	44.14	32.69	84.10	32.94	41.05	35.63
TCN (Bai et al., 2018)				82.55	31.84	42.83	32.44	82.22	29.82	38.01	34.96
Transformer (Vaswani et al., 2017)				82.29	31.84	43.10	31.46	83.04	31.79	40.02	35.69
Transformer (Vaswani et al., 2017)		√		85.27	35.44	45.84	35.69	85.87	35.45	42.85	39.27
Deepr (Nguyen et al., 2017)				81.05	28.56	38.81	31.45	82.90	30.59	28.97	34.09
Agent (Gao et al., 2020a)				83.84	34.06	45.68	32.51	82.50	30.98	39.42	34.65
AdaCare (Ma et al., 2020a)				79.48	27.30	40.04	26.52	81.08	29.20	38.39	30.28
ConCare (Ma et al., 2020b)				79.62	27.54	41.27	24.50	81.33	30.04	38.93	31.83
GRASP (Zhang et al., 2021)				79.70	27.63	40.73	24.90	81.44	29.81	38.68	31.56
RETAIN (Choi et al., 2016b)				83.95	34.55	46.57	33.03	83.34	31.84	40.06	35.71
StageNet (Gao et al., 2020b)				82.10	31.01	42.74	30.79	82.36	31.19	39.45	33.45
DeepSeek-zero shot (DeepSeek-AI, 2025)				53.13	10.60	28.55	13.27	54.75	10.59	20.87	15.77
DeepSeek-few shot (DeepSeek-AI, 2025)				55.62	16.20	30.74	20.20	56.40	14.09	23.70	21.05
MedTok (Su et al., 2025)	√			82.53	30.07	41.54	28.34	79.06	23.92	33.10	27.56
TRANS (Chen et al., 2024)	√		√	81.78	31.52	43.56	30.90	/	/	/	/
MultEHR (Chan et al., 2024)			√	81.98	31.43	42.65	**40.58**	81.60	30.27	39.16	35.39
GraphCare (Jiang et al., 2024)	√		√	80.50	27.90	43.50	36.80	83.00	32.50	40.60	35.90
HEAL-Net				83.17	32.66	45.06	31.80	83.97	33.06	41.50	34.90
		√		86.50	37.23	48.95	35.32	86.73	37.84	45.27	39.97
			√	83.81	33.21	45.38	32.20	84.05	34.03	42.14	36.10
		√	√	86.74	**38.32**	49.22	36.69	86.68	**38.15**	**45.62**	**40.32**
	√			84.01	34.67	46.00	35.62	83.79	32.73	41.20	33.85
	√	√		86.25	36.50	47.86	34.72	86.85	37.28	44.97	38.66
	√	√	√	**86.98**	38.25	**49.60**	36.17	**87.51**	37.93	45.61	39.42

fact that ordinal calibration—which AUROC, κ, and Accuracy reward and macro-F1 only weakly captures—is precisely what the Markovian `dir_G` head and the severity-aware LAB branch are designed to deliver.

Drug Recommendation. HEAL-Net achieves the best sample-F1 and sample-Jaccard on both datasets (68.49/54.15 on MIMIC-III, 66.33/50.21 on MIMIC-IV), and is competitive on the probabilistic metrics. DemoGraph attains the nominally highest sample-AUPRC/AUROC but collapses on F1/Jaccard (≤ 13 in both datasets), reflecting a well-known calibration pathology where probabilities rank well but fall below the deployment threshold; HEAL-Net is the only model that is

Table 3: Experimental Results for Drug Recommendation (Task 4).

Model	Input			MIMIC-III				MIMIC-IV			
	TS	Lab	Graph	AUPRC	AUROC	F1-score	Jaccard	AUPRC	AUROC	F1-score	Jaccard
CNN (LeCun et al., 1998)				66.70	90.45	58.01	42.08	73.98	94.13	61.76	46.63
CNN (LeCun et al., 1998)		√		70.13	92.86	63.95	47.78	77.54	95.26	66.22	**51.69**
RNN (Elman, 1990)				74.65	93.01	63.46	47.81	74.69	94.35	62.69	47.58
TCN (Bai et al., 2018)				73.60	92.52	62.66	46.88	73.09	93.77	60.82	45.69
Transformer (Vaswani et al., 2017)				72.80	92.38	61.93	46.03	70.45	93.05	58.39	43.13
Transformer (Vaswani et al., 2017)		√		73.10	92.93	62.61	46.81	72.96	93.91	61.40	46.15
Deepr (Nguyen et al., 2017)				73.64	92.70	62.27	46.36	72.73	93.67	60.37	45.08
MoleRec (Yang et al., 2023b)				45.93	75.15	58.94	43.21	42.15	72.50	54.88	39.51
Agent (Gao et al., 2020a)				74.38	92.66	63.20	47.44	73.71	94.02	61.56	46.43
AdaCare (Ma et al., 2020a)				62.96	**96.19**	11.91	46.97	62.49	97.68	11.06	47.22
ConCare (Ma et al., 2020b)				60.04	95.57	10.89	44.22	59.61	97.43	9.61	45.20
GRASP (Zhang et al., 2021)				62.39	95.98	11.65	46.22	62.79	**97.70**	11.13	47.06
RETAIN (Choi et al., 2016b)				74.41	92.92	63.04	47.27	74.98	94.44	62.72	47.66
GAMENet (Shang et al., 2019b)				71.52	92.05	61.26	45.49	71.65	93.16	59.63	44.96
GAMENet (Shang et al., 2019b)		√		72.21	92.11	61.97	46.12	72.64	93.78	60.58	45.36
StageNet (Gao et al., 2020b)				73.27	92.79	61.89	46.25	73.69	93.97	61.19	46.18
DeepSeek-zero shot (DeepSeek-AI, 2025)				31.55	62.36	35.57	22.89	30.07	64.89	38.16	25.43
DeepSeek-few shot (DeepSeek-AI, 2025)				41.00	72.57	53.12	37.33	39.19	74.10	52.44	37.35
MICRON (Yang et al., 2021)				73.10	92.93	62.61	46.81	75.07	93.78	60.58	45.36
MedTok (Su et al., 2025)	√			67.42	90.46	57.00	40.77	70.79	93.32	59.35	44.06
TRANS (Chen et al., 2024)	√		√	74.81	93.19	63.03	47.31	/	/	/	/
MultEHR (Chan et al., 2024)			√	73.13	96.04	55.09	43.71	66.66	96.57	46.88	36.25
GraphCare (Jiang et al., 2024)	√		√	78.90	94.30	65.20	48.70	77.40	95.80	64.00	49.20
HEAL-Net				74.62	93.11	63.20	47.61	72.78	93.79	60.08	44.86
		√		78.38	94.95	67.49	53.04	76.81	94.36	64.52	48.46
			√	76.05	93.52	64.86	49.40	73.47	94.60	62.19	45.83
		√	√	79.28	95.23	68.00	53.69	77.42	96.10	65.38	48.83
	√			74.36	92.92	63.35	47.71	73.12	94.25	60.57	45.20
	√	√		77.57	94.93	67.50	52.91	78.20	95.89	65.17	49.19
	√	√	√	**79.39**	95.25	**68.49**	**54.15**	**78.63**	96.10	**66.33**	50.21

simultaneously strong on probabilistic and threshold-based metrics, which we view as the relevant criterion for clinically deployable prescribing. Adding TS does not help on this task (consistent with prescribing being driven by current diagnoses and prior medications rather than physiological waveforms), and the target-visit DRUG drop guarantees that ATC-3 labels do not leak through the hash-code embedding.

Overall. A single backbone—differing only in the task-adaptive routing block and fusion mode—reaches the top of every task. This is the central empirical claim we set out to test: a temporally unrolled, relation-aware HEAL-Graph furnishes a unified inductive bias that transfers across acute, ordinal, and combinatorial clinical objectives without per-task architectural redesign.

4.2 Ablation

We perform three controlled ablations to validate the three core design choices of HEAL-Net: (i) the graph topology, (ii) the cross-visit encoding scheme, and (iii) the laboratory feature representation. In each study we vary only the targeted component while keeping all remaining inputs, training schedules, and routing blocks identical. Results are reported on MIMIC-III with one principal metric per task (Mortality AUROC, Readmission AUROC, LOS Cohen's κ, Drug sample-F1).

4.2.1 Graph Topology: Simple Co-occurrence Graph vs. HEAL-Graph

We first contrast our temporally unrolled, heterogeneous, relation-typed HEAL-Graph with a *simple* co-occurrence graph that aggregates all medical concepts into a single visit-level homogeneous adjacency without temporal unrolling, sub-relational typing, or knowledge-augmented edges. As reported in Table 4, replacing HEAL-Graph with the simple graph leads to substantial degradation across all four tasks. The collapse on multi-label drug recommendation is particularly diagnostic: when all visits share a flat homogeneous adjacency, current-visit DRUG nodes can no longer be cleanly separated from historical ones, so message passing over the simple graph leaks signal across labels and destroys the threshold-based prediction. By contrast, HEAL-Graph's per-visit node instantiation, sub-relational typing, and target-visit DRUG drop preserve a clean inductive prior that simultaneously benefits acute, ordinal, and combinatorial tasks.

Table 4: Controlled ablations on MIMIC-III. Each variant changes only the named component of HEAL-Net; the full model is the shared reference.

Variant	Mort. AUROC	Readm. AUROC	LOS κ	Drug F1
Simple co-occurrence graph	71.57	65.27	36.45	51.44
Direct concatenation (raw codes + raw lab)	72.41	69.18	37.21	65.47
Raw continuous lab values	72.14	70.18	36.72	64.59
HEAL-Net (full model)	**73.69**	**71.11**	**38.25**	**68.49**

4.2.2 Cross-Visit Encoding: Direct Concatenation vs. Cross-Visit Encoder

We next examine the contribution of HEAL-Net's cross-visit encoder. The baseline replaces the encoder with naive direct concatenation: per-visit raw codes and lab values are flattened across all historical visits and fed into the downstream head, which is the simplest multi-visit fusion strategy and corresponds to a *bag-of-visits* treatment. Table 4 shows that direct concatenation underperforms our cross-visit encoder by -1.28 Mortality AUROC, -1.93 Readmission AUROC, -1.04 LOS κ, and -3.02 Drug F1. The largest gap appears on the two longitudinal tasks (Mortality and Readmission), confirming that recency-aware aggregation, time-decay attention, and concept-equivalence linkage—all absent from a flattened bag-of-visits—are essential for resolving acute deterioration and re-admission risk. The persistent gap on LOS and Drug further indicates that cross-visit encoding is beneficial even when the immediate visit dominates the predictive signal.

4.2.3 Laboratory Feature Design: Raw Continuous Values vs. Severity-Discretized Encoding

Finally, we investigate the impact of the laboratory feature design. The baseline embeds raw continuous lab values directly as LAB node features without normality filtering, while our method discretizes each measurement into a small set of severity buckets (*normal* / mild / moderate / severe abnormal) and removes LAB nodes whose values fall in the normal range, so that only clinically actionable deviations participate in message passing. As shown in Table 4, the severity-discretized encoding improves all four tasks. The largest improvement is observed on Mortality, where in-hospital deterioration is most directly driven by abnormal physiology and noise from in-range measurements is most harmful; the smaller but consistent gains on the remaining tasks confirm that abnormality-filtered, severity-aware lab features improve signal-to-noise across heterogeneous clinical objectives without task-specific tuning.

Together, the three ablations isolate the contribution of each design axis and consistently support our central modelling claims: temporally unrolled heterogeneous topology, encoder-based cross-visit aggregation, and severity-aware lab encoding are all necessary for HEAL-Net's overall performance, and each removes a distinct failure mode of prior simpler designs.

5 Conclusion

We presented HEAL-Net, a heterogeneous evidence-adaptive framework for clinical prediction with unevenly available patient histories. HEAL-Net combines visit-aware clinical graphs, evidence-aware relational attention, and a shared temporal kernel with task-adaptive heads for binary, multi-label, and ordinal outcomes. Experiments on MIMIC-III and MIMIC-IV demonstrate strong performance across four clinical tasks, while ablations support the contributions of graph structure, temporal encoding, and severity-aware laboratory features. These results highlight the value of selective evidence integration and task-compatible historical adaptation. Future work will examine generalization across healthcare settings and robustness to missing or imperfect evidence.

AI USE STATEMENT

During the preparation of this manuscript, we used generative AI tools to assist with language polishing, including improving grammar, sentence structure, and the clarity of technical descriptions. These tools also supported the organization of the LaTeX source and the adaptation of manuscript formatting to the conference template. The authors reviewed the AI-assisted revisions for accuracy and consistency with the intended scientific meaning, paying particular attention to the presentation of the proposed method and the interpretation of experimental results. All decisions about the inclusion and wording of the revised material remained with the authors. We take full responsibility for the final manuscript, including its technical claims, reported results, and conclusions.

REPRODUCIBILITY STATEMENT

Our code is available in the anonymous repository at `https://anonymous.4open.science/r/med-58C1`. Appendix B describes the datasets, preprocessing, prediction tasks, evaluation metrics, and computing environment. Appendix C illustrates the graph construction process, and Appendix D provides dataset statistics and empirical analyses to support reproducibility.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. doi: 10.1016/j.patrec.2020.11.008.

Tsai Hor Chan, Guosheng Yin, Kyongtae Bae, and Lequan Yu. Multi-task heterogeneous graph learning on electronic health records. *Neural Networks*, 180:106644, 2024. doi: 10.1016/j.neunet.2024.106644.

Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang. Predictive modeling with temporal graphical representation on electronic health records. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5763–5771, 2024. doi: 10.24963/ijcai.2024/637.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: predicting clinical events via recurrent neural networks. In Finale Doshi-Velez, Jim Fackler, David C. Kale, Byron C. Wallace, and Jenna Wiens (eds.), *Proceedings of the 1st Machine Learning for Healthcare, MLHC 2016, Los Angeles, CA, USA, August 19-20, 2016*, JMLR Workshop and Conference Proceedings, pp. 301–318. JMLR.org, 2016a.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3504–3512, 2016b.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 787–795, 2017. doi: 10.1145/3097983.3098126.

Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 4552–4562, 2018.

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556, 2025. doi: 10.48550/ARXIV.2512.02556. URL https://doi.org/10.48550/arXiv.2512.02556.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. doi: 10.1207/s15516709cog1402_1.

Yushi Feng, Tsai Hor Chan, Guosheng Yin, and Lequan Yu. Democratizing large language model-based graph data augmentation via latent knowledge graphs, 2025. URL https://arxiv.org/abs/2502.13555.

Junyi Gao, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Dr. agent: Clinical predictive model via mimicked second opinions. *Journal of the American Medical Informatics Association*, 27(7):1084–1091, 2020a. doi: 10.1093/jamia/ocaa074.

Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M. Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (eds.), *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pp. 530–540. ACM / IW3C2, 2020b.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *CoRR*, abs/1703.07771, 2017.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *The Web Conference*, 2020.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha A. Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 05 2016. doi: 10.1038/sdata.2016.35.

Alistair Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei Lehman, Leo Celi, and Roger Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 01 2023. doi: 10.1038/s41597-022-01899-x.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaïne, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi Khorshidi. BEHRT: transformer for electronic health records. *CoRR*, abs/1907.09538, 2019.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4228–4238, 2021.

Chang Lu, Tian Han, and Yue Ning. Context-aware health event prediction via transition functions on dynamic disease graphs. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 4567–4574. AAAI Press, 2022.

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, 2020a. doi: 10.1609/aaai.v34i01.5427.

Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 833–840, 2020b. doi: 10.1609/aaai.v34i01.5428.

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. `Deepr`: A convolutional net for medical records. *IEEE J. Biomed. Health Informatics*, 21(1):22–30, 2017.

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. doi: 10.1038/s41746-021-00455-y.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 5953–5959, 2019a. doi: 10.24963/ijcai.2019/825.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 1126–1133. AAAI Press, 2019b.

Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Sahneh, and Marinka Zitnik. Multimodal medical code tokenizer. *International Conference on Machine Learning*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. Change matters: Medication change prediction with recurrent residual networks. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 3728–3734. ijcai.org, 2021.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P. Danek, and Jimeng Sun. Pyhealth: A deep learning toolkit for healthcare applications. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (eds.), *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pp. 5788–5789. ACM, 2023a.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben (eds.), *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pp. 4075–4085. ACM, 2023b.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. Medpath: Augmenting health risk prediction via medical knowledge paths. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (eds.), *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pp. 1397–1409. ACM / IW3C2, 2021.

Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. GRASP: Generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 715–723, 2021. doi: 10.1609/aaai.v35i1.16152.

Appendices

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Methodology **3**

3.1 Problem Formulation 3

3.2 Heterogeneous Clinical Graph 4

3.3 Relational Graph Attention 5

3.4 Temporal Visit Attention 5

3.5 Task-Adaptive Prediction 6

4 Experiments **6**

4.1 Main Results 6

4.2 Ablation 8

4.2.1 Graph Topology: Simple Co-occurrence Graph vs. HEAL-Graph 8

4.2.2 Cross-Visit Encoding: Direct Concatenation vs. Cross-Visit Encoder 9

4.2.3 Laboratory Feature Design: Raw Continuous Values vs. Severity-Discretized Encoding 9

5 Conclusion **9**

A Limitations and Future Work **16**

B Detailed Experimental Setup **16**

C Detailed Graph Construction **17**

C.1 Case Study: Graph Construction for Multi-Visit Patients 17

D Dataset Statistics and Empirical Analysis **18**

D.1 Dataset Overview 18

D.2 Label Distribution 18

D.3 Visit Complexity 18

D.4 Inter-Visit Time Gaps 20

D.5 Code Recurrence 20

D.6 Inter-Visit Jaccard Similarity 21

D.7 LOS Transition Structure 21

D.8 Information-Theoretic Analysis 21

D.9 Vocabulary and Code Frequency 23

D.10 Padding and Sample Composition 23

D.11 LOS Trajectory Types 23
D.12 Summary of Empirical Findings 24

Table 5: MIMIC-III and MIMIC-IV Dataset Statistics

	#patients	#visits	#visits/ patient	#conditio- ns/patient	#procedu- res/patient	#drugs/ patient
mimic3-mortality	6,186	9,717	1.5708	12.1146	3.6057	48.4532
mimic3-readmission	6,186	9,717	1.5708	12.1146	3.6057	48.4532
mimic3-length of stay	35,707	44,399	1.2434	10.3925	3.649	43.9621
mimic3-drug recommendation	5,449	14,141	2.5952	31.652	9.4601	29.7427
mimic4-mortality	71,086	160,711	2.2608	13.2578	2.5254	41.6448
mimic4-readmission	71,086	160,711	2.2608	13.2578	2.5254	41.6448
mimic4-length of stay	145,836	278,064	1.9067	12.4811	2.614	41.4461
mimic4-drug recommendation	55,446	187,674	3.3848	51.8442	9.0274	24.607

A Limitations and Future Work

While HEAL-Net provides a robust mechanism for structured multi-visit EHR modeling, it relies primarily on discrete codes and structured laboratory values. Future extensions of this work will explore the integration of unstructured clinical narratives (e.g., discharge summaries and nursing notes) directly into the graph topology using large language models. Additionally, while our temporal decay functions effectively handle irregular sampling intervals, evaluating the framework's resilience to systemic missingness and differing data imputation strategies across diverse healthcare systems remains an important area for future investigation. Ultimately, we envision HEAL-Net serving as a versatile and interpretable foundational architecture for next-generation clinical decision support systems.

B Detailed Experimental Setup

Datasets. We evaluate HEAL-Net on two widely used public EHR benchmarks: **MIMIC-III** (v1.4) (Johnson et al., 2016) and **MIMIC-IV** (v3.1) (Johnson et al., 2023). MIMIC-III contains 58,976 hospital admissions for 46,520 adult patients, and MIMIC-IV scales this to 546,028 admissions for 223,452 patients, with both datasets providing diagnoses, procedures, prescriptions, and laboratory measurements. We adopt the standard PyHealth (Yang et al., 2023a) preprocessing pipeline: each patient is organized as a chronologically ordered sequence of visits, and each visit is associated with diagnosis (ICD-9/ICD-10), procedure (ICD-9/ICD-10-PCS), drug (ATC), and lab features. For MIMIC-III we use the `PATIENTS`, `ADMISSIONS`, `DIAGNOSES_ICD`, `PROCEDURES_ICD`, `PRESCRIPTIONS`, and `LABEVENTS` tables; for MIMIC-IV we use the corresponding `patients`, `admissions`, `diagnoses_icd`, `procedures_icd`, `prescriptions`, and `labevents` tables. Per-task cohort statistics are summarized in Table 5. The roughly 10–20$\times$ scale gap between MIMIC-III and MIMIC-IV further provides a stress test of methodological scalability.

Tasks. We evaluate HEAL-Net on four clinically meaningful prediction tasks: (1) **Mortality prediction**, a binary classification task predicting whether a patient will die during or shortly after the current hospitalization; (2) **Readmission prediction**, a binary classification task predicting whether a patient will be re-admitted within 15 days after discharge; (3) **Length-of-stay (LOS) prediction**, a 10-class classification task that maps the length of the current stay into 10 ordered duration bins (with class 0 representing <1 day and class 9 representing >14 days); and (4) **Drug recommendation**, a multi-label classification task that predicts the set of medications (defined over the ATC level-3 vocabulary) prescribed during the current visit. The four tasks span heterogeneous label structures (binary, ordinal multi-class, multi-label) and signal sources (acute physiology, longitudinal therapeutic trajectory, diagnostic structure), constituting a stringent test of representational versatility.

Baselines. We compare HEAL-Net against 17 representative baselines covering five categories. *Standard temporal/spatial encoders*: CNN, RNN, TCN, and Transformer (Vaswani et al., 2017). *Specialized clinical sequence models*: Deepr (Nguyen et al., 2017), RETAIN (Choi et al., 2016b), StageNet (Choi et al., 2016b), Agent, and MICRON (Yang et al., 2021). *Drug-recommendation specialists*: GAMENet (Shang et al., 2019b) and MoleRec (Yang et al., 2023b). *LLM-based clinical predictors*: DeepSeek (DeepSeek-AI, 2025) (zero-shot and few-shot) and KARE (Jiang et al., 2025). *Graph- and knowledge-augmented EHR models*: DemoGraph (Feng et al., 2025) and Graph-

Care (Jiang et al., 2024). GraphCare in particular represents the strongest knowledge-graph EHR baseline currently available; we therefore use it as our primary point of comparison among graph-based methods. All baselines are run with their official PyHealth or upstream implementations and recommended hyperparameters. To reuse their published encoders without distortion, models that natively consume only event codes (e.g., RNN, TCN, Agent, RETAIN, GAMENet) operate on the discrete code streams, while models that admit auxiliary inputs (CNN, Transformer, Deepr, GAMENet) are additionally augmented with our laboratory representation, indicated by a check in the *Lab* column. The check-mark columns *TS*/*Lab*/*Graph* in Tables 1–3 indicate which input modalities each row consumes: TS denotes the raw temporal event stream, Lab denotes the severity-discretized laboratory channel, and Graph denotes the HEAL-Graph branch.

Evaluation Metrics. Because the four tasks have heterogeneous label structures, we use a metric set tailored to each task. All metrics are computed via `sklearn.metrics` to align with prior work. For binary tasks (Mortality, Readmission), we report **AUPRC** (`average_precision_score`, the stepwise area under the precision–recall curve) and **AUROC** (`roc_auc_score`, equivalent to the Mann–Whitney U rank statistic). For the multi-class LOS task, we report macro one-vs-rest **AUROC** (per-class binary AUROC averaged uniformly over the 10 classes), **Cohen's** κ ($\kappa = (p_o - p_e)/(1 - p_e)$ with p_o the observed agreement and p_e the chance agreement under the marginal distribution), **Accuracy**, and **macro-F1**. Hard predictions are obtained by $\arg\max_k p_{i,k}$. For the multi-label drug-recommendation task, all metrics are sample-averaged: sample-AUROC and sample-AUPRC compute a per-sample binary curve over the K drug labels and then average over samples (samples with only positive or only negative labels are skipped, following the `sklearn` convention), while sample-F1 and sample-Jaccard apply a 0.3 probability threshold and compute $2|L_i \cap \hat{L}_i|/(|L_i| + |\hat{L}_i|)$ and $|L_i \cap \hat{L}_i|/|L_i \cup \hat{L}_i|$ per sample, averaged over samples.

Machines and GPUs. Our experiments were conducted on a workstation featuring dual 64-core AMD EPYC 7742 CPUs and 512GB RAM, integrated with eight NVIDIA A100 (40GB) GPUs for hardware acceleration.

C Detailed Graph Construction

C.1 Case Study: Graph Construction for Multi-Visit Patients

To demonstrate the robustness of the HEAL-Graph generation process, we detail the graph instantiation for a real patient (MIMIC-III Subject ID 36) exhibiting highly variable temporal dynamics across three distinct admissions. This trajectory perfectly highlights the framework's ability to natively filter edges and contextualize diagnoses.

Patient Clinical Trajectory Summary:

- **Visit 1 (V1):** An elective coronary artery bypass graft (CABG) surgery (LOS: 8 days). Primary drivers are Coronary Artery Disease (CAD, `CCSCM_101`), COPD (`CCSCM_127`), and associated medications (Aspirin, Atorvastatin).
- **Visit 2 (V2):** Occurring a mere 12 days post-discharge, this admission is characterized by acute post-operative complications, specifically Pulmonary Embolism (PE, `CCSCM_103`), Deep Vein Thrombosis (DVT, `CCSCM_118`), and operative wound disruption.
- **Visit 3 (V3):** Occurring approximately 1080 days ($\approx$ 3 years) after V2. The primary cause is a ventral hernia (`CCSCM_143`) complicated by hospital-acquired pneumonia (`CCSCM_122`) and an abnormally high INR lab result indicating warfarin toxicity.

Graph Construction Dynamics Analysis:

- **Temporal Edge Filtering (Short vs. Long Intervals):** The HEAL-Graph successfully links V1 to V2 utilizing robust `evolve` edges, due to the extremely short interval (12 days), well within the continuous temporal decay parameter $\tau_d = 180$. Conversely, the transition from V2 to V3 spans roughly 1080 days. Consequently, standard evolutionary connections inherently face a strict cutoff, accurately reflecting the dissociation of the acute V2 events from the distal V3 events.

Table 6: Cohort sizes for MIMIC-III and MIMIC-IV.

Statistic	MIMIC-III	MIMIC-IV
# Patients	35,707	145,536
# Visits	44,399	277,396
Visits / patient	1.24	1.91
Multi-visit (≥ 2) patients	5,449 (15.3%)	55,329 (38.0%)
Single-visit patients	30,258 (84.7%)	90,207 (62.0%)
# Consecutive visit pairs	8,692	131,860

Table 7: Train/Val/Test split (visit-level samples, 8:1:1, seed 42).

	MIMIC-III		MIMIC-IV	
Split	# Samples	%	# Samples	%
Train	35,584	80.1	221,685	79.9
Val	4,398	9.9	28,043	10.1
Test	4,417	9.9	27,668	10.0
Total	44,399	100.0	277,396	100.0

- **Chronic Continuity vs. Acute Divergence:** Chronic conditions such as CAD (`CCSCM_101`) and COPD (`CCSCM_127`) are structurally maintained across all three visits, establishing continuous pathways via `evolve` and `same_concept` bridging. PE (`CCSCM_103`) correctly recurred in V2 and V3, generating strong recurrent semantic ties. The framework natively isolates acute pathologies—such as the V1 CABG, V2 wound dehiscence, and V3 pneumonia—allowing the integration of localized `intra_clinical` relationships without bleeding irrelevant historical information into current acute subgraphs.

D Dataset Statistics and Empirical Analysis

This appendix reports descriptive statistics, visit-level complexity, code recurrence, inter-visit similarity, label transition structure, and information-theoretic analyses for the two Length-of-Stay (LOS) datasets used in our experiments: MIMIC-III and MIMIC-IV. Unless otherwise noted, all statistics are computed on the cleaned cohorts after PyHealth preprocessing. LOS labels follow the standard 10-class scheme (classes 0–7 correspond to days 0–7, class 8 to 8–14 days, and class 9 to >14 days).

D.1 Dataset Overview

Table 6 summarizes the cohort sizes. MIMIC-III contains only 15.3% multi-visit patients, whereas MIMIC-IV provides much richer longitudinal history with 38.0% multi-visit patients. Table 7 reports the train/val/test split (8:1:1, seed 42).

D.2 Label Distribution

Both datasets exhibit class imbalance (Table 8). MIMIC-III is heavily skewed toward long stays: classes 8 and 9 jointly account for 46.8% of visits, with empirical entropy $H(Y)$=2.0913 nats (max $\log 10 \approx 2.3026$). MIMIC-IV is more balanced ($H(Y)$=2.2401 nats). Stratifying by visit count shows that the label distribution of multi-visit patients is nearly identical to the overall distribution (entropy difference <0.02 nats), indicating that history length itself does not introduce a label shift.

D.3 Visit Complexity

Table 9 reports per-visit code-count statistics across the four feature types (Conditions, Procedures, Drugs (ATC3 for MIMIC-III; raw NDC-derived for MIMIC-IV), Labs). Table 10 shows that the number of codes per visit is strongly positively correlated with LOS in both datasets, with Drugs and Labs being the strongest signals.

Table 8: Overall LOS class distribution (%). $H(Y)$ in nats; max entropy $\log 10 \approx 2.3026$.

Class	Meaning	MIMIC-III	MIMIC-IV
0	$<$1 day	2.8	6.2
1	1 day	4.2	13.3
2	2 days	6.6	14.6
3	3 days	7.4	12.1
4	4 days	9.2	10.9
5	5 days	8.4	7.9
6	6 days	7.6	6.1
7	7 days	7.0	4.9
8	8–14 days	25.9	14.7
9	$>$14 days	20.9	9.1
$H(Y)$ (nats)	—	2.0913	2.2401

Table 9: Per-visit code-count statistics.

Feature	Dataset	Mean	Median	P5	P25	P75	P95	Max
Conditions	III	10.4	9	4	7	14	21	34
	IV	12.1	11	3	7	16	25	46
Procedures	III	3.6	3	1	2	5	9	26
	IV	2.6	2	1	1	3	7	34
Drugs	III	44.0	35	5	23	48	151	246
	IV	41.4	25	9	17	40	153	293
Labs	III	68.6	67	30	48	86	117	247
	IV	46.0	41	0	24	64	105	276

Table 10: Spearman correlation ρ between #codes per visit and LOS class.

Feature	MIMIC-III	MIMIC-IV
Conditions	0.341	0.527
Procedures	0.523	0.272
Drugs	0.526	0.688
Labs	0.487	0.662

Table 11: Distribution of `delta_days` between consecutive visits.

Statistic	MIMIC-III	MIMIC-IV
# Pairs	8,692	131,860
Mean (days)	364.9	455.2
Median	115.0	118.0
Std	556.7	735.2
P25 / P75	26 / 454	24 / 570
P95 / P99	1,611 / 2,523	2,078 / 3,474
Max	3,487	5,346

Table 12: Binned distribution of `delta_days` (%).

Gap	MIMIC-III	MIMIC-IV
0–7 d	10.0	8.9
7–30 d	17.1	19.3
30–90 d	18.8	17.5
90–180 d	12.4	11.0
180–365 d	12.5	11.1
365–730 d	12.4	11.6
$>$730 d	16.7	20.7

D.4 Inter-Visit Time Gaps

We compute `delta_days` between consecutive visits of the same patient. Table 11 reports summary statistics; Table 12 reports binned distributions. Both datasets cover an extremely wide range (up to $\sim$10 years for MIMIC-III and $\sim$15 years for MIMIC-IV). About 90% of consecutive visits are separated by ≥ 7 days, suggesting that consecutive visits are almost always independent admissions rather than continuations of a single hospitalization.

D.5 Code Recurrence

Define the *recurring fraction* as the fraction of codes in the current visit that have appeared in *any* prior visit of the same patient. Table 13 reports recurrence statistics on multi-visit samples; Table 14 correlates recurrence with LOS.

The pattern is consistent across datasets: *higher recurrence* (i.e., the patient's current visit largely revisits known conditions/medications/labs) is associated with *shorter LOS*, with Labs and Drugs giving the strongest negative correlation. Procedures exhibit a bimodal recurrence distribution (concentrated near 0 or near 1), consistent with episodic surgery vs. chronic recurring procedures. Drugs and Labs are heavily right-skewed, reflecting routine medications and standing lab orders.

Table 13: Recurring fraction (multi-visit samples only).

Feature	Dataset	Mean	Median	Mean #new	Mean #total
Conditions	III	0.529	0.500	6.1	12.9
	IV	0.470	0.500	7.2	13.8
Procedures	III	0.445	0.400	2.1	3.5
	IV	0.324	0.000	1.8	2.4
Drugs	III	0.681	0.714	10.9	30.3
	IV	0.725	0.789	8.2	25.4
Labs	III	0.791	0.830	16.1	69.1
	IV	0.748	0.838	14.5	49.6

Table 14: Spearman ρ between recurring fraction and LOS class.

Feature	MIMIC-III	MIMIC-IV
Conditions	-0.209	-0.036
Procedures	-0.140	-0.008
Drugs	-0.209	-0.147
Labs	-0.278	-0.265

Table 15: Mean Jaccard: same-patient consecutive visits vs. random different-patient pairs.

Feature	Dataset	Same-pt.	Diff.-pt.	Ratio
Conditions	III	0.325	0.088	$3.7\times$
	IV	0.267	0.025	$10.7\times$
Procedures	III	0.228	0.079	$2.9\times$
	IV	0.180	0.008	$21.7\times$
Drugs	III	0.428	0.278	$1.5\times$
	IV	0.434	0.217	$2.0\times$
Labs	III	0.543	0.463	$1.2\times$
	IV	0.452	0.338	$1.3\times$

D.6 Inter-Visit Jaccard Similarity

Table 15 compares the Jaccard similarity of code sets between consecutive visits of the *same* patient against random pairs from *different* patients. Table 16 reports the Spearman correlation between Jaccard and time gap.

Same-patient consecutive visits are markedly more similar than random pairs, confirming that patient-level history carries a structured, learnable signal. The decay with time is mild: even after >1 year, mean Conditions Jaccard remains around 0.17–0.27.

D.7 LOS Transition Structure

We treat *Copy-last*—predicting that the current LOS class equals the previous one—as a non-trivial baseline. Table 17 shows that this baseline is bounded above by ~ 18–21% exact accuracy, far below the strong code-driven baselines reported in our main experiments. Table 18 reports the mean per-class change $\Delta = c_{\text{cur}} - c_{\text{prev}}$, which exhibits a strong regression-to-the-mean pattern in both datasets.

D.8 Information-Theoretic Analysis

Table 19 reports mutual information (MI) in nats between several signals and the current LOS, together with the ratio $\mathrm{MI}/H(Y_{\text{cur}})$. The shuffle-corrected MI of the previous LOS is essentially zero in both datasets: *the value of the previous LOS itself carries virtually no information about the current LOS*. In contrast, code counts of the current visit carry 5–$10\times$ more information; the strongest signals are #Drugs and #Labs in MIMIC-IV ($\mathrm{MI}/H \approx 13\%$).

Table 16: Spearman ρ between Jaccard similarity and inter-visit gap.

Feature	MIMIC-III	MIMIC-IV
Conditions	-0.279	-0.405
Procedures	-0.174	-0.231
Drugs	-0.282	-0.205
Labs	-0.188	-0.216

Table 17: Copy-last baseline statistics on consecutive visit pairs.

Metric	MIMIC-III	MIMIC-IV
Exact match (%)	21.1	18.0
Within ± 1 class (%)	45.8	41.6
Mean abs. error (classes)	2.45	2.63
Spearman($c_{\text{prev}}, c_{\text{cur}}$)	0.203	0.260

Table 18: Mean LOS-class change Δ given starting class.

Starting class	MIMIC-III	MIMIC-IV
0	$+5.45$	$+3.45$
1	$+3.96$	$+2.77$
2	$+3.21$	$+1.79$
3	$+2.28$	$+1.14$
4	$+1.71$	$+0.44$
5	$+0.87$	-0.21
6	$+0.18$	-1.04
7	-0.97	-1.91
8	-1.59	-2.62
9	-2.27	-3.08

Table 19: Mutual information $\mathrm{MI}(X; Y_{\text{cur}})$ in nats and $\mathrm{MI}/H(Y_{\text{cur}})$.

	MIMIC-III		MIMIC-IV	
Signal X	MI	MI/H (%)	MI	MI/H (%)
Previous LOS (raw)	0.0301	1.45	0.0480	2.14
Previous LOS (shuffle-corr.)	**0.0000**	0.00	**0.0000**	0.00
#Conditions (binned)	0.0680	3.25	0.1629	7.27
#Procedures (binned)	0.1571	7.51	0.0694	3.10
#Drugs (binned)	0.2146	10.26	0.3018	13.47
#Labs (binned)	0.1430	6.84	0.2832	12.64
ΔLOS (trend)	0.0058	0.28	0.0014	0.06

Table 20: Vocabulary size, head coverage, and singleton ratio.

Feature	Dataset	$\lvert\mathcal{V}\rvert$	# Occurrences	Top-10% cov.	Top-25% cov.	Singletons
Conditions	III	281	461,417	50.6%	79.3%	1 (0.4%)
	IV	8,774	3,361,819	90.1%	97.5%	1,428 (16.3%)
Procedures	III	221	162,013	68.6%	87.3%	5 (2.3%)
	IV	2,751	725,015	81.9%	94.9%	309 (11.2%)
Drugs	III	197	1,951,872	52.5%	81.0%	1 (0.5%)
	IV	201	11,495,176	53.4%	80.3%	0 (0.0%)
Labs	III	706	3,047,470	72.3%	95.9%	26 (3.7%)
	IV	896	12,747,069	81.6%	95.9%	16 (1.8%)

Table 21: Distribution of real visit counts per training sample.

Real visits	III #	III %	IV #	IV %
1	28,565	80.3	116,428	52.5
2	4,359	12.2	44,263	20.0
3	1,314	3.7	21,159	9.5
4	579	1.6	11,762	5.3
5	767	2.2	28,073	12.7

D.9 Vocabulary and Code Frequency

Table 20 reports vocabulary sizes and head/tail concentration. After CCS/ATC3 grouping in MIMIC-III, vocabulary sizes are small (≤ 706) with negligible long tail. MIMIC-IV uses the original ICD/ItemID codes, leading to substantially larger vocabularies for Conditions and Procedures and a non-trivial singleton fraction (11–16%).

D.10 Padding and Sample Composition

Sequences are left-padded to a maximum history length of $M{=}5$ visits; position 4 is always the current visit. Table 21 reports the distribution of true visit counts in the training set; Table 22 reports the fill rate at each padded position. Overall padded-slot fractions are 73.4% for MIMIC-III and 58.9% for MIMIC-IV. We verify that all padded slots have `delta_days`$=0$, so the alignment is exact.

Because over 80% of MIMIC-III training samples have only one real visit, models cannot use sequence length $T{>}1$ as a proxy for "has history"; instead, the actual code content of each padded slot must be inspected.

D.11 LOS Trajectory Types

Table 23 categorizes the LOS trajectory of multi-visit patients. Only 11–16% of patients have a perfectly stable LOS class across visits, providing further evidence that the prior LOS value cannot be naively extrapolated.

Table 22: Per-position fill rate (fraction of samples with real data).

Position	MIMIC-III	MIMIC-IV
pos 0	2.2%	12.7%
pos 1	3.8%	18.0%
pos 2	7.5%	27.5%
pos 3	19.7%	47.5%
pos 4	100.0%	100.0%

Table 23: LOS trajectory types among multi-visit patients (%).

Trajectory	MIMIC-III	MIMIC-IV
Always decreasing	32.7	23.1
Always increasing	26.3	24.1
Fluctuating	25.3	41.7
Perfectly stable	15.7	11.2

Table 24: Summary of empirical findings and their design implications.

Finding	Evidence	Implication
Previous LOS value carries almost no information	Shuffle-corrected MI$=0$	Do not use prior LOS as a direct prior
Copy-last accuracy is low	18%–21% exact match	"Replicate-history" methods are inherently bounded
Single-visit samples dominate	84.7% (III), 62.0% (IV)	Models must be robust without history
#codes per visit is a strong LOS signal	ρ up to 0.69	Prioritize current-visit code content
Recurrence anti-correlates with LOS	$\rho \approx -0.20$ to -0.28	Known-disease management is faster than new workups
Jaccard decays mildly with time	$\rho \approx -0.20$ to -0.40	Long-range history retains residual signal
Strong regression to the mean	Δ flips sign at class $\sim$ 5–6	Trend itself carries little predictive information
Current-visit MI 5–10$\times$ that of history	0.30 vs. 0.03 nats (IV-Drugs)	History should be auxiliary, not primary
MIMIC-IV has long-tailed vocabulary	Singleton ratio 16.3% for Conditions	Embedding regularization or grouping is needed
Padding dominates	73.4% (III), 58.9% (IV)	Use code-content masks rather than length

D.12 Summary of Empirical Findings

Table 24 summarizes the empirical observations and their implications for model design.