Hierarchical Modeling of ICD Codes in EHR Foundation Models

Megha Thukral[1] MTHUKRAL3@GATECH.EDU
Dong Gyun Kang[1] DKANG335@GATECH.EDU
Rudra Pratap Singh[1] RUDRA.SINGH@GATECH.EDU
Shruthi Kashinath Hiremath[2] SHRUTHI.HIREMATH@OPTUM.COM
Katrin Hänsel[2] KATRIN_HAENSEL@OPTUM.COM
Thomas Plötz[1] THOMAS.PLOETZ@GATECH.EDU

[1] *School of Interactive Computing, Georgia Institute of Technology, Atlanta, United States*
[2] *Optum AI, United States*

Abstract

Electronic health record foundation models typically treat ICD diagnosis codes as flat tokens, overlooking the clinically meaningful hierarchical structure that captures disease families, subcategories, and fine-grained diagnostic detail. As a result, existing EHR representation learning methods do not explicitly exploit the hierarchical structure already present in the coding system. In this work, we study ICD-10-CM hierarchy as a *general inductive bias* for clinical representation learning. We investigate two complementary mechanisms for incorporating hierarchy: first, by augmenting diagnosis sequences in a BERT-style transformer with tokens corresponding to different levels of the ICD hierarchy, and second, by injecting hierarchy into graph-based code representations through hierarchy-aware edges combined with diagnosis co-occurrence structure. Across these settings, we evaluate whether explicit hierarchy improves downstream prediction, which levels of the hierarchy are most useful, whether hierarchy encoding improves transfer across datasets, and how hierarchy reshapes embedding similarity structure. We conduct experiments on two large-scale real-world clinical datasets: MIMIC-IV, used for pretraining and four in-domain prediction tasks, and eICU, used to assess cross-dataset transfer via frozen encoder probing. In our experiments, we find that explicitly encoding ICD hierarchy improves over flat code representations in both in-domain and cross-dataset settings, while revealing that the most useful level of hierarchy depends on both the task and the modeling approach. Broadly, we focus on hierarchy-aware EHR representation learning and show that the benefits of encoding hierarchy are generalizable across modeling settings and hierarchy levels.

1. Introduction

Electronic health records (EHRs) provide a longitudinal view of patients' clinical history through diagnoses, procedures, medications, and encounters. Advances in transformer-based EHR models, such as BEHRT and Med-BERT, have shown that large-scale pretraining can learn useful patient representations for downstream clinical prediction tasks (Li et al., 2020; Rasmy et al., 2021). However, much of this progress still treats structured

clinical codes as flat symbols, even when those codes are drawn from ontologies that were explicitly designed to encode clinical hierarchy. This creates a mismatch between the structure of the data and the structure presented to the model. For instance, ICD codes are not arbitrary identifiers: they organize diseases into clinically meaningful families, subgroups, and increasingly specific categories (World Health Organization, 2016) (see Figure 1). A code, therefore, carries both a fine-grained diagnosis and a path through a broader clinical taxonomy. In practice, however, most EHR models represent ICD codes as independent tokens, ignoring this hierarchical structure. As a result, related diagnoses that should share information may instead be learned as isolated symbols. This is especially problematic in clinical data, where many diagnoses are sparse, long-tailed, and institution-specific (Choi et al., 2016a).

We hypothesize that explicitly modelling hierarchy could benefit EHR representation learning in several ways. First, the hierarchy provides a natural inductive bias for relationships between diseases, so rare diagnosis codes can benefit from data associated with clinically similar codes (Perotte et al., 2014). Second, broader and intermediate groupings may capture care pathways or anatomical systems that are more predictive than the most specific code alone (Song et al., 2019). Third, hierarchy can improve robustness under *domain shift*, since higher-level disease families are often more stable across institutions (Ostrominski et al., 2021). More broadly, encoding hierarchy can help learned embeddings better reflect clinical semantics rather than only empirical co-occurrence.

Prior work has explored incorporating ICD code hierarchy into healthcare representation learning, such as GRAM (Choi et al., 2017), primarily through graph-based approaches that embed ontological structure into code representations. However, these methods predate the current generation of transformer-based EHR foundation models, and it remains underexplored how the ICD hierarchy should be systematically incorporated into such models, and which granularity of hierarchy proves most beneficial.

To address this gap, we propose and evaluate two complementary strategies for injecting ICD hierarchy into contemporary EHR representation learning. The first, `HICD-BERT` (**H**ierarchical **ICD**-BERT), injects hierarchy directly into token representations in a BERT-style encoder, embedding each level of the ICD hierarchy as an additive token embedding alongside the diagnosis code. The second, `HICD-Graph` (**H**ierarchical **ICD**-Graph), encodes hierarchy relationally via a diagnosis co-occurrence graph augmented with ontology-derived edges, learning hierarchy-aware code embeddings that initialise a patient-level Transformer. Together, these approaches allow us to compare whether hierarchical information is best consumed as a token-level signal or as a relational structure over the code vocabulary.

Our results show that hierarchy is a *robustly effective inductive bias* for EHR representation learning, as across both architectures and all four prediction tasks, the large majority of hierarchy-augmented configurations outperform their no-hierarchy baselines. Both models benefit from hierarchy, with graph-based encoding leveraging all hierarchy levels and token-based encoding benefiting most from the finest granularity level. Crucially, hierarchy also improves cross-dataset transfer, specifically for graph-based hierarchy encoding, which transfers robustly from MIMIC-IV to eICU.

- We investigate ICD code hierarchy as an inductive bias for EHR representation learning across two architectures, four clinical prediction tasks, and a systematic ablation

over three granularity levels, showing that hierarchy significantly improves performance in 50 of 56 comparisons.

- We examine the benefit of hierarchy under distributional shift by evaluating cross-dataset transfer from MIMIC-IV to eICU under a frozen-encoder probe protocol, showing that graph-based hierarchy encoding transfers robustly to new datasets and task settings.
- Through embedding analysis of the learned code representations, we show that hierarchy produces more coherent code clusters, and that the configurations with the tightest clusters also achieve the strongest downstream performance.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our findings suggest two broader lessons for machine learning in health. First, structured clinical ontologies such as ICD provide useful inductive biases for EHR foundation models, and even lightweight ways of encoding that structure can yield consistent gains over flat code representations. Second, incorporating clinically meaningful structure can also help under distribution shift, and the mechanism of incorporation can impact: graph-based hierarchy encoding transfers more robustly across datasets than token-level hierarchy injection. These insights are relevant beyond ICD-10-CM and can be extended to other structured clinical vocabularies, such as ATC and procedure ontologies, and to other representation learning methods.

2. Related Work

2.1. Sequential and Transformer-based EHR representation learning

Early work on longitudinal EHR modeling established the importance of learning from temporally ordered patient histories. Doctor AI (Choi et al., 2016a) used recurrent neural networks to predict future clinical events from visit sequences, while RETAIN (Choi et al., 2016c) introduced a reverse-time attention mechanism to improve interpretability in healthcare prediction. Dipole (Ma et al., 2017) further demonstrated that attention-based sequence models can improve diagnosis prediction by capturing dependencies across visits. These methods showed that sequential structure in EHRs is highly informative, but they generally treated diagnosis codes as flat symbols.

More recently, Transformer-based models have become a central approach for structured EHR representation learning. BEHRT (Li et al., 2020) was an early and influential framework that represents patient histories as sequences of medical codes and applies self-attention to capture dependencies across visits. Med-BERT (Rasmy et al., 2021) further showed that masked pretraining on large EHR corpora can produce reusable representations for downstream prediction tasks. Subsequent models such as CEHR-BERT (Pang et al., 2021), ExBEHRT (Rupp et al., 2023), and larger-scale foundation models such as Foresight (Kraljevic et al., 2024) and ETHOS (Renc et al., 2024) have further scaled this paradigm with richer input representations and larger training corpora.

Generative transformers have further shown that this paradigm scales to whole-life disease forecasting and massive real-world datasets: Delphi-2M (Shmatko et al., 2025) models

the progression and competing risk of over 1,000 diseases across an individual's lifespan, validated cross-nationally on UK Biobank and Danish registry data, while Waxler et al. (Waxler et al., 2025) establish power-law scaling relationships for medical event modeling using 118 million patients from Epic Cosmos. Despite these advances, Transformer-based EHR models have generally treated diagnosis codes as atomic identifiers.

2.2. Ontology-aware and Hierarchy-aware Medical Representation Learning

A parallel line of work has explored how medical ontologies can improve representation learning. Many of these methods were developed in the pre-foundation-model era and are trained end-to-end for a single downstream task rather than as reusable pretrained encoders. GRAM (Choi et al., 2017) incorporates ontology ancestors via attention-weighted aggregation within a recurrent patient encoder, and introduced an intuition central to our study: clinically related diagnoses should share information through their position in a hierarchy. KAME (Ma et al., 2018) extended this with a knowledge-based attention mechanism for diagnosis prediction. MiME (Choi et al., 2018) explored related ideas by structuring medical codes around clinically meaningful groupings. Other works that use ICD hierarchies in classical predictive settings include propagated-binary rollups (Singh et al., 2014), hierarchy-level analyses for hospital-day prediction (Xie et al., 2015), and hierarchy-aware multi-task transformers for EHR classification (Blanco et al., 2021). Among ontology-aware methods that use pretraining, G-BERT (Shang et al., 2019) applies graph-augmented BERT-style pretraining for medication recommendation, though it does not study how different ontology levels contribute. Recently, Zhou and Barbieri (2025) incorporates hierarchy-aware structure for synthetic EHR generation.

A complementary line of work has explored purely data-driven approaches to code relationships, learning structure from co-occurrence rather than from an external taxonomy. Med2Vec (Choi et al., 2016b) demonstrated that skip-gram-style co-occurrence embeddings can recover clinically meaningful code and visit representations without any ontology supervision. Subsequent graph-based methods have modeled empirical dependencies among diagnoses for prediction: (Poulain and Beheshti, 2024) builds patient-level heterogeneous graphs that jointly encode diagnoses, procedures, and demographics, learning representations through message passing over clinically grounded relations, while (Xi et al., 2025) constructs disease co-occurrence graphs from patient records and applies graph neural networks to capture higher-order relational signal beyond pairwise embeddings. These approaches recover relational structure directly from data, but unlike ontology-aware methods, they do not leverage existing domain knowledge-based hierarchy.

Across these lines of work, prior methods typically rely on a single source of hierarchical signal. Even when pretrained, they generally do not ablate across hierarchy levels or explicitly examine how hierarchy itself shapes the learned representation. This leaves open whether hierarchy helps consistently across tasks and architectures, which levels of the ICD hierarchy contribute most, and how hierarchy-aware representations behave under cross-dataset transfer. In contrast, we evaluate two paradigms head-to-head within modern Transformer-based EHR foundation models: `HICD-BERT` injects hierarchy at the token-embedding level via string-prefix truncation, requiring no ontology access at training time, while `HICD-Graph` combines ontology-derived hierarchy edges with empirical PMI-weighted

co-occurrence edges in a hybrid diagnosis graph, learning hierarchy-aware code embeddings via a graph convolutional network. Across both paradigms, we systematically ablate three hierarchy granularity levels, evaluate on four in-domain tasks and one cross-dataset transfer task, and analyze how hierarchy reshapes the learned embedding geometry, isolating not only *whether* hierarchy helps, but *which* encoding paradigm makes it most useful, *which* levels contribute, and *whether* ontology information is strictly necessary.

3. Methods

We represent each patient as a temporally ordered sequence of visits, $\mathcal{V} = (v_1, \ldots, v_T)$, where each visit v_t contains one or more ICD diagnosis codes. Our goal is to learn patient representations that preserve both temporal context and the hierarchical structure of diagnosis ontologies, and to evaluate whether such structure improves downstream prediction and generalization. We study two complementary mechanisms for injecting ICD hierarchy into EHR models, which we collectively refer to as **H**ierarchical **ICD** models (**HICD**). The first is a *token-level* approach, `HICD-BERT`, which extends BEHRT (Li et al., 2020) by adding hierarchy-specific embedding tokens for each diagnosis code (Figure 2(a)). The second is a *graph-level* approach, `HICD-Graph`, which constructs a hierarchy-augmented diagnosis co-occurrence graph, learns code embeddings with a graph convolutional network (GCN) (Kipf and Welling, 2017), and uses those embeddings to initialise a Transformer-based patient encoder (Figure 2(b)). Full implementation details and hyperparameters are provided in Appendix B.

3.1. ICD Hierarchy Levels

ICD-10-CM codes are organised hierarchically (World Health Organization, 2016). Every code belongs to a *chapter* (a broad body-system or etiology grouping, e.g. Chapter XIX, *Injury, Poisoning and External Causes*), partitioned into officially defined *blocks* of related categories (e.g. `S70--S79`, *Injuries to the hip and thigh*), which contain three-character *categories* (e.g. `S72`, *Fracture of femur*). Characters after the decimal may encode etiology, anatomic site, severity, and encounter details (Fig. 1). Each leaf code is nested inside progressively coarser clinical groupings, and this nesting is the inductive bias we inject.

For each diagnosis code, we define three nested hierarchy levels `G0`, `G1`, `G2`, ordered from coarsest to finest. These levels are relative-granularity indices, each architecture operationalises them through its own mechanism (details below). Each ablation setting enables or disables a subset of these levels, yielding all eight combinations $(\texttt{G0}, \texttt{G1}, \texttt{G2}) \in \{0,1\}^3$, including the non-hierarchical baseline $(0,0,0)$, equivalent to BEHRT (Li et al., 2020) for `HICD-BERT` and to a standard GCN-Transformer without hierarchy edges for `HICD-Graph`. This setup compares two paradigms: data-driven token encoding, where hierarchy is derived directly from diagnosis code

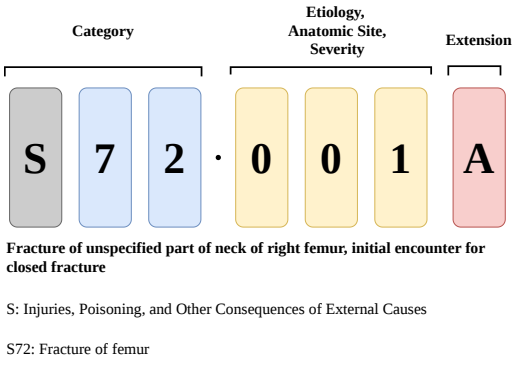


Figure 1: Structure of an ICD-10-CM diagnosis code.

strings, and hybrid graph-based encoding, where hierarchy is injected through ontology-derived graph nodes alongside empirical co-occurrence structure.

`HICD-BERT` **(data-driven).** Hierarchy is injected as auxiliary token embeddings using string-prefix truncation (Fig. 2(a)), requiring no external ontology at training time. `G0` is the 1-character alphabetic prefix, `G1` the 2-character prefix, and `G2` the 3-character ICD category. These act as data-derived surrogates for increasing hierarchy depth rather than official chapter or block identifiers. For example, `S72.001A` maps to $(\texttt{G0}, \texttt{G1}, \texttt{G2}) = (\texttt{S},\ \texttt{S7},\ \texttt{S72})$.

`HICD-Graph` **(hybrid ontology-based).** Hierarchy is represented as graph nodes connected to leaf diagnosis codes by weighted edges (Fig. 2(b)). These nodes are derived from the ICD-10-CM ontology and are incorporated into a graph that also encodes diagnosis co-occurrence structure. `G0` denotes the ICD-10 chapter, `G1` the official ICD-10-CM block, and `G2` the 2-character prefix group. For example, `S72.001A` maps to $(\texttt{G0}, \texttt{G1}, \texttt{G2}) =$ (Chapter XIX, S70–S79, S7x). Although `G1` and `G2` are closely related, they are not identical: `G1` follows official ICD-10-CM block boundaries, while `G2` groups codes by shared first two characters. A worked example of how leaf codes share hierarchy nodes in `HICD-Graph` is provided in Appendix A. All analyses and comparisons are made *within* each model, and we interpret the levels by relative granularity (coarse to fine) rather than assuming the same label corresponds to identical clinical content across architectures.

3.2. `HICD-BERT`: Hierarchy-aware BEHRT

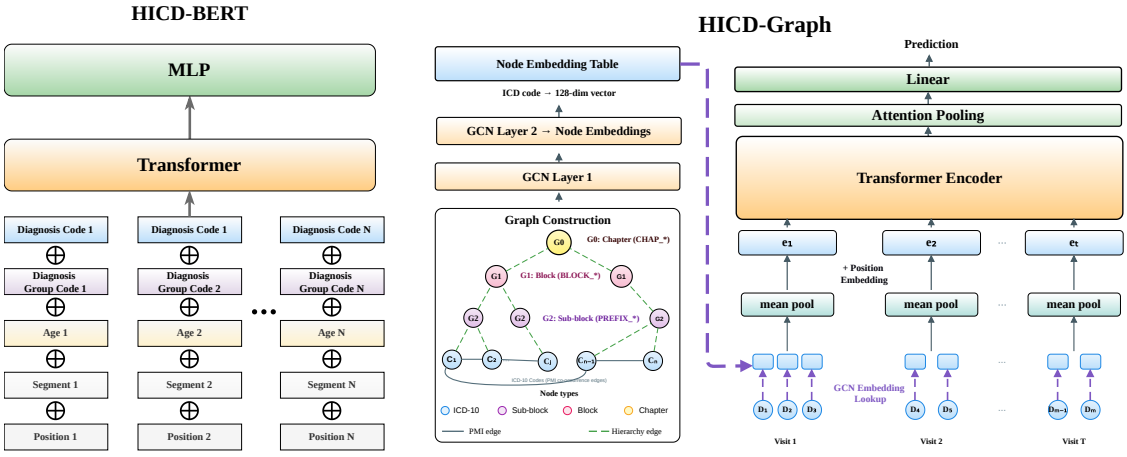
`HICD-BERT` extends BEHRT (Li et al., 2020), a Transformer encoder for longitudinal EHR sequences. In BEHRT, each diagnosis token is represented by the sum of the token, age, positional, and segment embeddings. We augment this representation with up to three hierarchy-specific embeddings:

$$\mathbf{h}_i^{(0)} = \mathbf{e}_{\text{code}}(c_i)+\mathbf{e}_{\text{age}}(a_i)+\mathbf{e}_{\text{pos}}(p_i)+\mathbf{e}_{\text{seg}}(s_i)+\mathbb{I}[G0]\mathbf{e}_{G0}(g_i^{(0)})+\mathbb{I}[G1]\mathbf{e}_{G1}(g_i^{(1)})+\mathbb{I}[G2]\mathbf{e}_{G2}(g_i^{(2)}), \tag{1}$$

where c_i is the diagnosis token, a_i is age, p_i is position, s_i is the segment identifier, and $g_i^{(0)}, g_i^{(1)}, g_i^{(2)}$ are the group identifiers at hierarchy levels $G0$, $G1$, and $G2$. This design preserves the original BEHRT tokenization while directly injecting the coarse-to-fine ICD structure into the embedding layer. The resulting sequence is processed by a stack of Transformer encoder blocks (Vaswani et al., 2017). Each block contains multi-head self-attention, a position-wise feed-forward network, residual connections, and layer normalization. The hierarchy-aware embedding layer is the only change to the BEHRT backbone. Additional architectural and training details are provided in Appendix B.1. We pretrain `HICD-BERT` using a masked language modeling (MLM) objective over the diagnosis sequence. Given a masked sequence $\tilde{\mathbf{x}}$, the model predicts the original diagnosis token at masked positions:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log p(c_i \mid \tilde{\mathbf{x}}), \tag{2}$$

where $\mathcal{M}$ denotes the set of masked positions. For downstream prediction, we initialize the encoder from the pretrained MLM checkpoint and attach a task-specific multilayer perceptron to the final `[CLS]` representation. Binary tasks are optimized with binary cross-entropy with logits, while multiclass tasks are optimized with cross-entropy.



(*a*) HICD-BERT (*b*) HICD-Graph

Figure 2: Comparison of HICD-BERT and HICD-Graph.

3.3. `HICD-Graph`: Graph-based Hierarchy Modeling

Our second model injects ICD hierarchy at the code-graph level rather than the token-embedding level. We first construct a diagnosis co-occurrence graph over ICD codes, where horizontal edges reflect empirical co-occurrence within patient visits and are weighted by pointwise mutual information (PMI):

$$\mathrm{PMI}(a,b) = \log \frac{P(a,b)}{P(a)P(b)} + \epsilon, \tag{3}$$

where $P(a)$ and $P(b)$ are marginal code frequencies and $P(a,b)$ is their joint co-occurrence probability. We retain only positive PMI edges above a percentile threshold to suppress weak associations. We augment this empirical graph with uniform-weight ICD-10 hierarchy edges linking each diagnosis node to up to three enabled ancestor levels, `G2`, `G1`, and `G0` (Section 3.1, Fig. 4), yielding a hybrid graph construction that combines data-driven PMI edges with ontological hierarchy edges. Given the resulting graph, we learn code embeddings using a two-layer GCN (Kipf and Welling, 2017) trained with a link-prediction objective:

$$\mathcal{L}_{\text{graph}} = -\sum_{(u,v)\in\mathcal{E}} \log \sigma(\mathbf{z}_u^\top \mathbf{z}_v) - \sum_{(u,v^-)\notin\mathcal{E}} \log\left(1 - \sigma(\mathbf{z}_u^\top \mathbf{z}_{v^-})\right), \tag{4}$$

where $\mathbf{z}_u$ and $\mathbf{z}_v$ denote node embeddings and $\mathcal{E}$ is the set of graph edges. The learned code embeddings are then used to initialise the embedding table of a patient-level Transformer encoder. For each visit, we compute a visit-level representation by applying masked mean pooling over the embeddings of the diagnosis codes (and, when enabled, their hierarchy ancestor codes) observed in that visit. The temporally ordered sequence of visit representations is then processed by a Transformer encoder, followed by attention-based pooling across visits to produce a patient-level representation, which is passed to a task-specific classification head. Additional implementation details are provided in Appendix B.2.

3.4. Experimental Setup

3.4.1. Datasets

MIMIC-IV We use data from the Medical Information Mart for Intensive Care IV (MIMIC-IV), a large, de-identified clinical database developed at Beth Israel Deaconess Medical Center in Boston, Massachusetts (Johnson et al., 2023). MIMIC-IV contains electronic health records for hospital admissions between 2008 and 2019. For this work, we restrict our use of MIMIC-IV to *diagnosis information* from the hospital-wide module. This module covers all inpatient admissions, not only ICU stays, and stores structured diagnosis codes assigned to each hospital encounter. Diagnoses are encoded using International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)(mapped to ICD-10 in our work) and Tenth Revision, Clinical Modification (ICD-10-CM). Each diagnosis code is linked to dictionary tables providing a textual description and hierarchical groupings, which we leverage to define clinical conditions and construct label spaces.

eICU Collaborative Research Database The eICU Collaborative Research Database (eICU-CRD) is a multi-center critical care database containing de-identified health data for over 200,000 ICU admissions from more than 200 hospitals across the United States (Pollard et al., 2018). Data were collected through the Philips eICU telehealth program and span a wide range of ICU types, including medical, surgical, cardiac, and mixed units. eICU-CRD provides diagnosis codes recorded using ICD-9-CM and ICD-10-CM, as well as proprietary diagnosis fields used within the telehealth platform.

Both MIMIC-IV and eICU-CRD are de-identified, hosted on PhysioNet (Goldberger et al., 2000), and available to credentialed researchers who complete the required data use training and agreements.

3.4.2. Task Definitions

We evaluate all models on four binary classification tasks defined over patient visit sequences from MIMIC-IV, and one binary classification task on eICU for transfer evaluation. For all MIMIC-IV tasks, the final visit of each patient is excluded from the prediction set, as no subsequent outcome label can be observed. Given a patient's visit history up to and including visit v_t, we define the following tasks.

30-Day Readmission (MIMIC-IV). The model predicts whether the patient will be readmitted within 30 days of discharge. The binary label is defined as

$$y_t^{\text{readmit}} = \mathbb{I}[\text{admittime}(v_{t+1}) - \text{dischtime}(v_t) < 30 \text{ days}]\,. \tag{5}$$

This task serves as a standard clinical benchmark for assessing whether hierarchy-aware representations capture patient risk signals that are predictive of near-term care utilization.

Emergency-Admission Prediction (MIMIC-IV). The model predicts whether the immediately following visit v_{t+1} is an emergency admission. MIMIC-IV records nine admission types; we define emergency admissions as the union of `URGENT`, `EW EMER.`, `EU OBSERVATION`, `OBSERVATION ADMIT`, `AMBULATORY OBSERVATION`, `DIRECT EMER.`, and `DIRECT OBSERVATION`, based on the visit type information provided by MIMIC-IV. The binary label

is

$$y_t^{\text{type}} = \mathbb{I}[\text{admission\_type}(v_{t+1}) \in \mathcal{T}_{\text{EM}}]\,, \tag{6}$$

where $\mathcal{T}_{\text{EM}}$ denotes the set of emergency admission types defined above. This task evaluates whether the learned representations encode temporal diagnostic patterns predictive of the urgency of future care.

Long Length-of-Stay (MIMIC-IV). The model predicts whether the next visit v_{t+1} will result in a long inpatient stay (> 7 days). The binary label is

$$y_t^{\text{LOS}} = \mathbb{I}[\text{dischtime}(v_{t+1}) - \text{admittime}(v_{t+1}) > 7 \text{ days}]\,. \tag{7}$$

This task probes whether hierarchy-aware representations capture markers of acuity and care intensity.

1-Year Post-Discharge Mortality (MIMIC-IV). The model predicts whether the patient will die within 365 days of the discharge of the next visit v_{t+1}. In-hospital deaths are excluded so that the label isolates post-discharge survival. The binary label is

$$y_t^{\text{mort}} = \mathbb{I}[0 \leq \text{dod} - \text{dischtime}(v_{t+1}) \leq 365 \text{ days}]\,, \tag{8}$$

where dod is the patient's date of death.

ICU Readmission Prediction (eICU). The eICU Collaborative Research Database organizes data at the level of individual ICU unit stays, with multiple unit stays possible within a single hospital admission, linked by a common `patientHealthSystemStayID`. Given the diagnosis codes observed during ICU unit stay u_k for a patient, the model predicts whether the patient will be readmitted to the ICU within the same hospital admission. The binary label is

$$y_k^{\text{readmit}} = \mathbb{I}[u_k \text{ is not the final ICU unit stay within the hospital admission}]\,. \tag{9}$$

This task is used exclusively for transfer evaluation: models are pretrained on MIMIC-IV and only a classifier head is fine-tuned on eICU, with the encoder kept frozen.

We split the data into an 80:10:10 ratio to create a train/validation/test split shared across all hierarchy ablations for a model. For `HICD-BERT`, we first pretrain the model with masked language modeling (ref. Sec. 3.2) and then fine-tune it for each downstream task. For `HICD-Graph`, we first learn hierarchy-aware code embeddings from the diagnosis graph and then use these embeddings to initialize the downstream Transformer-based patient encoder. In `HICD-BERT`, we evaluate all subsets of the three hierarchy levels `G0`, `G1`, and `G2`, which correspond to enabling or disabling hierarchy embeddings; in `HICD-Graph`, they correspond to enabling or disabling the associated hierarchy edges and ancestor nodes. We evaluate performance using F1, and AUROC scores. For transfer experiments, models trained on MIMIC-IV are evaluated on eICU without modifying the hierarchy construction pipeline. Exact optimization settings and hyperparameters are provided in Appendix B.

Table 1: 30-day readmission results for `HICD-BERT` and `HICD-Graph` across hierarchy ablations. Brackets show 95% bootstrap CIs. $^*p < 0.05$ vs. no hierarchy, using pairwise DeLong tests for AUROC and paired bootstrap tests for F1-macro.

Group	Method	`HICD-BERT`		`HICD-Graph`	
		F1-macro	AUROC	F1-macro	AUROC
Baselines	No hierarchy	0.5786 [.573,.584]	0.6807 [.674,.687]	0.5976 [.593,.602]	0.6992 [.694,.705]
	GRAM	0.5683 [.563,.574]	0.6825 [.677,.689]	0.6132 [.609,.618]	0.7120 [.707,.717]
	LR rollup	0.5049 [.500,.510]	0.6343 [.627,.641]	0.5784 [.574,.583]	0.6597 [.654,.666]
Single-level	G0	0.5825 [.577,.588]	0.6856* [.680,.692]	0.6025* [.598,.607]	0.7086* [.704,.714]
	G1	0.5740* [.569,.580]	0.6810 [.675,.687]	0.6006* [.596,.605]	0.7079* [.702,.713]
	G2	0.5896* [.584,.595]	0.6920* [.686,.698]	0.6005* [.596,.605]	0.7086* [.703,.714]
Two-Level	G0+G1	0.5696* [.564,.575]	0.6863* [.680,.692]	**0.6088*** [.604,.613]	0.7105* [.705,.716]
	G0+G2	**0.5953*** [.590,.601]	0.6918* [.686,.698]	0.5999 [.596,.604]	0.7097* [.704,.715]
	G1+G2	0.5834* [.578,.589]	0.6917* [.685,.698]	0.6000 [.596,.604]	0.7113* [.706,.717]
All-Levels	G0+G1+G2	0.5731* [.568,.578]	**0.6921*** [.686,.698]	0.5991 [.595,.603]	**0.7114*** [.706,.717]

4. Results

We report the in-domain downstream performance of hierarchy-aware modeling, cross-dataset transfer under a frozen-encoder probe setting, and embedding-level analyses. We evaluate all eight combinations of the three ICD-10 hierarchy levels (`G0`, `G1`, `G2`) across both `HICD-BERT` and `HICD-Graph`, resulting in `56` pairwise comparisons using *2 models* $\times$ *4 tasks* against the no-hierarchy baseline within each paradigm: BEHRT (Li et al., 2020) for `HICD-BERT` and a GCN + Transformer encoder for `HICD-Graph`. We add two hierarchy-aware baselines: GRAM (Choi et al., 2017), an ontology-attention model, and a propagated-binary ICD-10 rollup with regularised logistic regression (Singh et al., 2014). Full details in Appendix B.3. We report statistical significance using the paired DeLong test (DeLong et al., 1988) for all pairwise comparisons of binary AUROC on the same test set. For F1-macro, we use a paired bootstrap over 1,000 test-set resamples and a two-sided p-value, following prior work (Berg-Kirkpatrick et al., 2012).

4.1. 30-day Readmission Task

Table 1 shows that adding hierarchical structure improves 30-day readmission prediction for both `HICD-BERT` and `HICD-Graph` relative to the no-hierarchy baseline (DeLong $p < 0.05$ for 6 of 7 `HICD-BERT` variants and all 7 `HICD-Graph` variants).

For `HICD-BERT`, multi-level hierarchy outperforms single-level, with the finest level driving the gains. The strongest F1-macro is achieved by the two-level `G0+G2` configuration (0.5953), while the best AUROC is obtained when all hierarchy levels are enabled (0.6923). Among single-level configurations, the finest level `G2` yields the strongest improvement (AUROC 0.6921), while the mid-level `G1` is the only configuration that does not significantly improve over baseline, though it contributes meaningfully when combined with other levels. This suggests that for token-level hierarchy injection, fine-grained group-

Table 2: Emergency admission results for `HICD-BERT`and `HICD-Graph`across hierarchy ablations. Brackets show 95% bootstrap CIs. $^{*}p < 0.05$ vs. no hierarchy, using pairwise DeLong tests for AUROC and paired bootstrap tests for F1-macro.

Group	Method	`HICD-BERT`		`HICD-Graph`	
		F1-macro	AUROC	F1-macro	AUROC
Baselines	No hierarchy	0.4945 [.490,.500]	0.7682 [.760,.776]	0.6083 [.604,.613]	0.7495 [.745,.754]
	GRAM	0.5189 [.512,.526]	0.7818 [.774,.790]	0.6317 [.628,.636]	0.7568 [.753,.761]
	LR rollup	0.4886 [.484,.493]	0.7438 [.735,.752]	0.6104 [.606,.614]	0.7185 [.714,.722]
Single-level	G0	0.4846* [.481,.488]	0.7848* [.777,.792]	**0.6373*** [.633,.642]	0.7538* [.750,.758]
	G1	0.5107* [.504,.517]	0.7698 [.761,.778]	0.6209* [.616,.625]	0.7555* [.752,.760]
	G2	**0.5200*** [.513,.527]	**0.7941*** [.786,.802]	0.6269* [.623,.631]	0.7556* [.752,.760]
Two-Level	G0+G1	0.5138* [.507,.520]	0.7838* [.775,.791]	0.6343* [.630,.639]	0.7557* [.752,.760]
	G0+G2	0.5159* [.509,.523]	0.7924* [.784,.800]	0.6314* [.627,.636]	0.7557* [.752,.760]
	G1+G2	0.5117* [.505,.518]	0.7903* [.783,.798]	0.6313* [.627,.636]	**0.7577*** [.754,.762]
All-Levels	G0+G1+G2	0.4988 [.493,.504]	0.7917* [.784,.799]	0.6339* [.630,.638]	0.7567* [.752,.761]

ings are the primary driver of performance gains, and that mid-level groupings serve as connective structure between coarser and finer hierarchy rather than as standalone signal.

For `HICD-Graph`, more hierarchy levels yield progressively better readmission prediction. All seven hierarchy configurations yield statistically significant improvements over the no-hierarchy baseline, and a clear monotonic pattern emerges: single-level configurations improve AUROC by 1.3% on average over the baseline, two-level configurations by 1.6%, and the all-levels `G0+G1+G2` configuration by 1.7%. This progressive improvement suggests that each additional hierarchy level contributes complementary structure to the diagnosis graph, and that the graph-based architecture is able to effectively integrate information from multiple granularity levels simultaneously.

Taken together, the consistency of hierarchy benefits across two different pretraining objectives, a Transformer-based encoder pretrained with masked language modeling (`HICD-BERT`) and a GCN-initialized Transformer pretrained with link prediction on a diagnosis co-occurrence graph (`HICD-Graph`), supports that ICD hierarchy is a useful inductive bias for readmission prediction. While the absolute AUROC improvements are modest, they are statistically robust and architecturally consistent; we discuss the practical significance of these effect sizes in Section 5.

4.2. Emergency Admission Prediction

We report the results for emergency admission prediction in Table 2. Here again, we note that hierarchy improves over the no-hierarchy baseline, but the pattern of which configurations perform best differs from the readmission task.

For `HICD-BERT`, a single hierarchy level appears to be as effective as the full hierarchy. Unlike the readmission setting, where the all-levels configuration achieves the best AUROC, the single-level `G2` model performs as strongly as the all-levels variant on

Table 3: AUROC with 95% bootstrap CI for Long-LOS and 1-year mortality. * $p < 0.05$ vs. no-hierarchy baseline (paired DeLong tests). Full tables in Appendix B.10, B.11.

Task	Method	HICD-BERT	HICD-Graph
Long-LOS	No hierarchy	0.6727 [.665,.680]	0.6712 [.664,.679]
	LR Rollup (LR)	0.6298 [.623,.637]	0.6381 [.630,.646]
	GRAM	0.6752 [.669,.682]	0.6729 [.665,.681]
	HICD	**0.6840*** [.677,.691] (G1+G2)	**0.6840*** [.677,.691] (G0+G1+G2)
Mortality	No hierarchy	0.8181 [.812,.824]	0.7882 [.782,.795]
	LR Rollup	0.7685 [.762,.776]	0.7958 [.789,.803]
	GRAM	0.8269 [.821,.833]	0.7567 [.749,.764]
	HICD	**0.8356*** [.830,.842] (G0+G2)	**0.8275*** [.821,.834] (G0+G1+G2)

emergency-visit prediction, with both achieving comparable AUROC. As in readmission, the mid-level `G1` alone does not significantly improve over baseline, though it does yield a significant F1-macro gain. This task-dependent shift suggests that emergency-visit prediction can drive performance from a specific level of abstraction, and that adding coarser hierarchy levels on top of the finest grouping provides no additional benefit for this task.

For `HICD-Graph`, the monotonic trend from readmission holds. All seven variants significantly outperform the no-hierarchy baseline (all $p < 0.001$), and as in readmission, AUROC improves progressively from single-level to two-level to all-levels configurations. The best individual configuration is the two-level `G1+G2`, which achieves comparable or slightly better AUROC than the all-levels variant. This suggests that for emergency-visit prediction, the combination of intermediate and fine-grained hierarchy captures the most relevant diagnostic structure, and that adding the coarsest level (ICD chapter) on top provides no further benefit for this task.

4.3. Additional Clinical Outcomes: Long-LOS and 1-Year Mortality

Beyond near-term care utilization (readmission and emergency-admission), we additionally evaluate whether ICD hierarchy helps predict *clinical severity and longer-term outcomes*: long inpatient stays (> 7 days) and 1-year post-discharge mortality. These tasks have lower positive-class prevalence (≈15–19%) and provide a complementary test of whether hierarchy generalises beyond near-term utilization prediction. Table 3 summarises the no-hierarchy baseline against the best hierarchy variant and other baselines per architecture. The full eight-variant tables are deferred to Appendix B.10 (Long-LOS) and Appendix B.11 (Mortality). Hierarchy improves both tasks across both architectures, with the largest absolute gain observed in `HICD-Graph` on Mortality (+0.0393 AUROC). For `HICD-BERT`, the two-level `G1+G2` and `G0+G2` configurations dominate on Long-LOS and Mortality respectively, consistent with the finest level `G2` driving most of the single-level gain (as in the existing tasks). For `HICD-Graph`, the all-levels `G0+G1+G2` configuration is best for both tasks, mirroring the monotonic pattern seen in readmission. Both HICD variants outperform the two hierarchy-aware baselines (GRAM and the LR Rollup) for these tasks (Table 3). Across the four in-domain tasks, hierarchy-aware models outperform the no-hierarchy baseline in **50**

Table 4: Cross Dataset Transfer: Probe results on eICU readmission task for `HICD-BERT`and `HICD-Graph`across hierarchy ablations. $^*p < 0.05$ from pairwise DeLong tests on AUROC and paired bootstrap tests on F1-macro vs no hierarchy.

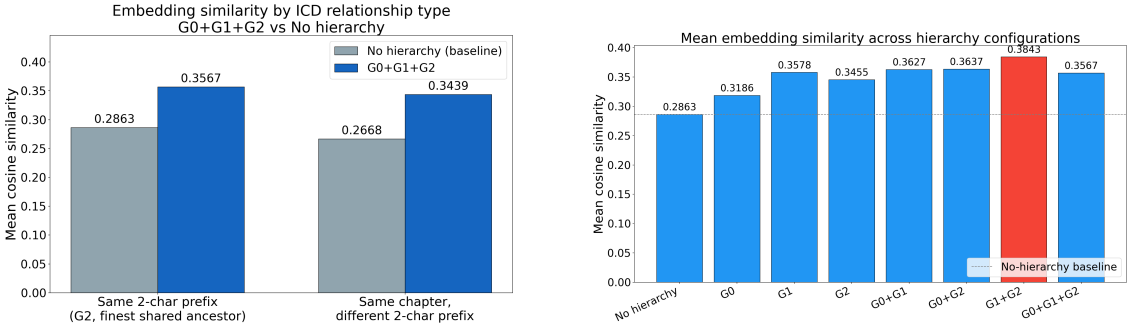
Group	Hierarchy encoding	`HICD-BERT`		`HICD-Graph`	
		F1-macro	AUROC	F1-macro	AUROC
Baseline	No hierarchy	0.4550 [.450,.460]	0.5653 [.554,.576]	0.3804 [.375,.385]	0.5645 [.553,.576]
Single-level	G0	0.4533 [.448,.458]	**0.5711*** [.565,.586]	0.3956* [.391,.401]	0.5757* [.565,.586]
	G1	0.4469* [.442,.452]	0.5641 [.552,.575]	**0.4866*** [.481,.492]	0.5783* [.567,.589]
	G2	**0.4848*** [.480,.490]	0.5650 [.554,.576]	0.4609* [.456,.466]	0.5832* [.572,.594]
Two-Level	G0+G1	0.4321* [.427,.437]	0.5682 [.557,.580]	0.4668* [.461,.472]	0.5819* [.570,.592]
	G0+G2	0.4332* [.428,.438]	0.5667 [.555,.578]	0.4871* [.482,.492]	0.5788* [.568,.590]
	G1+G2	0.4751* [.470,.480]	0.5683 [.556,.579]	0.4535* [.449,.458]	0.5738 [.563,.585]
All-Levels	G0+G1+G2	0.4743* [.469,.479]	0.5646 [.553,.576]	0.4743* [.469,.480]	**0.5871*** [.577,.597]

of 56 pairwise AUROC comparisons, demonstrating that ICD hierarchy provides a broadly effective inductive bias across tasks and architectures.

4.4. Cross-Dataset Transfer Results

Table 4 presents cross-dataset transfer results on eICU under a frozen-encoder probe protocol (Guo et al., 2024; Steinberg et al., 2024). In this setting, the encoder pretrained on MIMIC-IV is kept frozen, and only a task-specific classifier head is trained using eICU ICU-readmission labels. This evaluation measures whether the hierarchy-aware representations learned on one institution's data capture structure that generalises to a different patient population. We note that both `HICD-BERT` and `HICD-Graph` retain meaningful predictive signal under this setting, with all configurations outperforming chance-level prediction (AUROC > 0.55). However, the pattern of which hierarchy configurations benefit transfer differs between the two architectures and from the in-domain results.

Hierarchy benefits transfer more in `HICD-Graph` than in `HICD-BERT`. For `HICD-Graph`, 6 of 7 hierarchy variants significantly improve over the no-hierarchy baseline, with the all-levels `G0+G1+G2` configuration achieving the strongest AUROC (0.587). This demonstrates that the graph-based hierarchy encoding produces representations whose structure transfers robustly across datasets. In contrast, for `HICD-BERT`, only the coarsest single-level configuration `G0` improves AUROC over baseline (0.571); the remaining 6 variants are statistically indistinguishable from the no-hierarchy baseline, and F1-macro results are mixed, with only some variants improving F1 (`G2`, `G12`, `G0+G1+G2`). This is notable because `HICD-BERT` is the stronger beneficiary of hierarchy in the in-domain setting, where 6 of 7 variants achieve improvements, yet these gains do not survive the distributional shift to eICU. The asymmetry may reflect a difference in what the two encoders learn: the co-occurrence graph underlying `HICD-Graph` captures inter-code relationships that can be more institution-invariant, whereas the sequential token patterns learned by `HICD-BERT` are more tightly coupled to MIMIC-IV visit structures. Consistent with this interpretation, only the coarsest level `G0`



(*a*) Pairwise similarity. (*b*) Within-group similarity.

Figure 3: Embedding analysis for `HICD-Graph`. All hierarchy-aware variants increase semantic clustering over the baseline, with `G1+G2` achieving the strongest local clustering.

transfers for `HICD-BERT` as coarse groupings can be less susceptible to dataset-specific coding variation than finer-grained levels.

Hierarchy aware pretrained representations can capture transferable diagnosis structure. Despite the modest absolute AUROC values (0.56–0.59), these results supports that hierarchy-augmented pretraining produces more robust representations that can transfer well across datasets. The frozen encoder has not seen eICU patients or diagnoses during pretraining, yet hierarchy-aware configurations, particularly in `HICD-Graph`, achieve statistically significant improvements over the no-hierarchy baseline. This suggests that the ICD hierarchy provides an inductive bias that is not merely corpus-specific but reflects underlying relationships that can transfer well across datasets. To further strengthen transfer analysis, we report the relative MIMIC→eICU AUROC drop in Appendix B.6.

4.5. Embedding Analysis

We analyze the learned GCN node embeddings in `HICD-Graph` for leaf ICD-10 diagnosis codes to assess how hierarchical supervision changes the structure of the diagnosis-code representation space. We study pairwise cosine similarity under two ICD relationship types, defined by the shared ancestor in the ICD-10 ontology: codes sharing the same 2-character prefix (e.g. all codes beginning with `S7`), which corresponds to the model's `G2` grouping level, and codes from the same ICD-10 chapter but with different 2-character prefixes.

Hierarchy strengthens clinically meaningful similarity structure. Figure 3(a) compares the mean cosine similarity across these two relationship types for the no-hierarchy baseline and the `G0+G1+G2` configuration. Relative to the baseline, the `G0+G1+G2` configuration increases similarity in both groups, from 0.286 to 0.357 among codes sharing the same 2-character prefix, and from 0.267 to 0.344 among codes within the same chapter but with different prefixes. The resulting embedding space preserves a meaningful similarity gradient: codes sharing the finest ancestor (`G2`) are more similar than broader within-chapter codes,

supporting that hierarchical supervision strengthens the clinically relevant organization of the embedding space.

All hierarchy configurations improve local code similarity. Figure 3(b) examines how each hierarchy configuration affects pairwise similarity among diagnosis codes. All hierarchy-aware variants improve over the no-hierarchy baseline (0.286), confirming that explicit hierarchical information consistently pulls related diagnoses closer together in embedding space. The strongest effect is obtained when the mid-level and fine-level hierarchies are combined (`G1+G2`, 0.384), indicating that these two levels together capture the most clinically relevant grouping structure for organising related diagnosis codes in embedding space.

Embedding structure predicts downstream performance. Notably, these embedding-level patterns are also predictive of downstream task performance. On readmission prediction (Table 1), the `G1+G2` configuration, which produces the tightest local code clusters in embedding space, also achieves among the strongest AUROC for `HICD-Graph`, while the coarsest level `G0`, which yields the smallest embedding similarity gains, delivers the weakest single-level task performance. This alignment between representation geometry and predictive accuracy suggests that hierarchy improves `HICD-Graph` performance by shaping the embedding space in clinically meaningful ways that benefit downstream discrimination.

Hierarchy effects are heterogeneous across chapters in our pretraining data Figure 5 (Appendix) shows the intra-chapter cosine similarity for each ICD-10 chapter across all eight hierarchy configurations on our pretraining data, MIMIC-IV. Across the majority of chapters in our pretraining data, cosine similarity increases with the addition of hierarchy, confirming the global trend observed in Figure 3. However, the magnitude of improvement is heterogeneous: chapters with higher baseline similarity, such as Chapter XVIII ($0.48 \rightarrow 0.61$, $+27\%$) and Chapter XXII ($0.70 \rightarrow 0.90$, $+29\%$), exhibit comparatively modest relative gains, while lower-baseline chapters such as Chapter XVI ($0.09 \rightarrow 0.26$, $+189\%$) and Chapter XIX ($0.05 \rightarrow 0.10$, $+100\%$) show substantially larger proportional improvements under hierarchy. This suggests that hierarchical supervision is slightly more beneficial in chapters with greater code diversity, where the representation space is less coherent at baseline, and the model can be optimized more effectively by incorporating ontological structure. We provide analogous chapter-wise ΔAUROC tables and analysis in Appendix B.7.

5. Discussion

This work investigates ICD hierarchy as an inductive bias for EHR representation learning, evaluating two complementary mechanisms, token-level injection (`HICD-BERT`) and graph-level encoding (`HICD-Graph`). Our results support the view that the hierarchical structure already present in ICD coding systems is an *underutilised signal* in EHR foundation models, and that modeling it explicitly benefits downstream performance.

Key Findings. Across all four in-domain tasks, the finest hierarchy level `G2` is consistently the strongest single-level configuration for `HICD-BERT`, while `HICD-Graph` benefits progressively from adding more hierarchy levels. For `HICD-BERT`, the mid-level `G1` provides no significant improvement in isolation on either task, though it contributes when combined with other levels, suggesting that mid-level groupings serve as connective struc-

ture rather than a standalone signal. The optimal hierarchy depth is also task-dependent: readmission benefits from combining all levels, while emergency-visit prediction achieves comparable performance from a single fine-grained level alone. On cross-dataset transfer, hierarchy benefits `HICD-Graph` robustly (6 of 7 variants significant) but largely vanish for `HICD-BERT`, where only the coarsest level `G0` transfers. This suggests that graph-based relational structure, anchored in the ICD ontology, is more dataset-invariant than sequential token patterns. The embedding analysis provides mechanistic support: hierarchy encourages the embedding space to reflect clinical relationships, and the configurations producing the tightest local code clusters (`G1+G2`) also achieve among the strongest downstream AUROC, suggesting that the gains arise from meaningful representational structure.

Practical significance of effect sizes. Readmission, emergency-visit, length-of-stay, and mortality prediction are inherently difficult, multifactorial outcomes that depend on many factors beyond *diagnosis codes alone*. Against this backdrop, the AUROC improvements from hierarchy are statistically robust (50 of 56 comparisons significant) and consistent across two models and four tasks, supporting hierarchy as a general inductive bias. Importantly, in our setup, we incorporated hierarchy using only minimal modifications to existing architectures: additive token embeddings for `HICD-BERT` and additional graph edges for `HICD-Graph` with no changes to the core encoder, pretraining objective, or training pipeline. This *intentional low cost integration*, combined with the consistent improvements observed, suggests that the ICD hierarchy can be included as a strong prior in EHR representation learning.

Limitations. Our evaluation covers four binary clinical tasks on MIMIC-IV and one transfer task on eICU; the generalizability of our findings to other clinical outcomes, institutions, and EHR systems remains to be validated. Our analysis is restricted to ICD diagnosis codes; whether the observed benefits extend to other coded data in EHR foundational models remains an open question. The hierarchy levels (`G0`, `G1`, `G2`) are intentionally defined differently across `HICD-Graph` and `HICD-BERT` to evaluate two hierarchy construction strategies: ontology-derived grouping versus lightweight data-driven prefix grouping. All comparisons are therefore made within each model, while future work could analyze other encodings across architectures to isolate the effect of the hierarchy definition itself. Our hierarchy experiments are limited to a fixed set of manually defined levels, so we do not test whether the optimal hierarchy depth could be learned automatically for a given task or model.

Future Work. A natural next step is to extend our framework to multimodal EHR models and add hierarchy for other structured clinical vocabularies with analogous ontological structure, such as ATC codes for medications, which organize drugs into therapeutic, pharmacological, and chemical subgroups across multiple levels. We also plan to explore deeper integration mechanisms that go beyond the additive approaches studied here, including hierarchy-aware pretraining objectives that explicitly optimize for ontological coherence, adaptive mechanisms that learn to weight or select different hierarchy levels during training, and multi-ontology fusion that jointly encodes diagnosis, medication, and procedure hierarchies within a single model. We believe such approaches could further amplify the practical gains observed in this work.

Acknowledgments

This work was partially supported by a grant from Optum.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 995–1005, 2012.

Alberto Blanco, Alicia Perez, and Arantza Casillas. Exploiting icd hierarchy for classification of ehrs in Spanish through multi-task transformers. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1374–1383, 2021.

Centers for Medicare and Medicaid Services. General equivalence mappings (gems) for ICD-9-CM and ICD-10-CM/PCS. https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs, 2018. Accessed 2025.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016a.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504, 2016b. doi: 10.1145/2939672.2939823.

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, 2016c.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, 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*, pages 787–795, 2017.

Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. *Advances in neural information processing systems*, 31, 2018.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new

research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, June 2000. doi: 10.1161/01.CIR.101.23.e215.

Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. Adapting ehr foundation models across institutions via frozen-encoder probing. *npj Digital Medicine*, 2024.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL https://doi.org/10.1038/s41597-022-01899-x.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

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.

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*, pages 1903–1911, 2017.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 743–752, 2018.

John W Ostrominski, Javier Amione-Guerra, Brian Hernandez, Joel E Michalek, and Anand Prasad. Coding variation and adherence to methodological standards in cardiac research using the national inpatient sample. *Frontiers in cardiovascular medicine*, 8:713695, 2021.

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*, pages 239–260. PMLR, 2021.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237, 2014.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, Sep 2018. doi: 10.1038/sdata.2018.178.

Raphael Poulain and Rahmatollah Beheshti. Graph transformers on ehrs: Better representation improves downstream performance. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=pe0Vdv7rsL.

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.

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

Maurice Rupp, Oriane Peter, and Thirupathi Pattipaka. Exbehrt: Extended transformer for electronic health records. In *International workshop on trustworthy machine learning for healthcare*, pages 73–84. Springer, 2023.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*, 2019.

A. Shmatko, A. W. Jung, K. Gaurav, et al. Learning the natural history of human disease with generative transformers. *Nature*, 647:248–256, 2025. doi: 10.1038/s41586-025-09529-3. URL https://doi.org/10.1038/s41586-025-09529-3.

Anima Singh, Girish Nadkarni, John Guttag, and Erwin Bottinger. Leveraging hierarchy in medical codes for predictive modeling. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 96–103, 2014.

Lihong Song, Chin Wang Cheong, Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. Medical concept embedding with multiple ontological representations. In *IJCAI*, volume 19, pages 4613–4619, 2019.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. *ICLR*, 2024.

Xu Sun and Weichao Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale, 2025. URL https://arxiv.org/abs/2508.12104.

World Health Organization. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision*. World Health Organization, Geneva, 5th edition, 2016.

Suyang Xi, Jiesen Shi, Jiachen Yan, MingJing Lin, Xinyi Zhou, Yuan Cheng, Hong Ding, and Chia Chao Kang. Breaking barriers in icd classification with a robust graph neural network for hierarchical coding. *Scientific Reports*, 15(1):25676, 2025.

Yang Xie, Guenter Schreier, Mark Hoy, Ying Liu, Sandra Neubauer, David CW Chang, Stephen J Redmond, and Nigel H Lovell. Impact of hierarchies of clinical codes on predicting future days in hospital. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5728–5731. IEEE, 2015.

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

Appendix A. HICD-Graph Hierarchy: Example

Figure 4 illustrates how two leaf diagnosis codes share hierarchy nodes in `HICD-Graph`. Consider `I50.21` (*Acute systolic (congestive) heart failure*) and `I50.32` (*Chronic diastolic (congestive) heart failure*). Both leaves connect, via weighted edges, to the same `G2` node (`I5`x prefix group), the same `G1` node (official block `I30--I52`, *Other forms of heart disease*), and the same `G0` node (Chapter IX, *Diseases of the circulatory system*). Shared ancestors propagate information between clinically related leaves during message passing, while unrelated leaves remain separated in the hierarchy subgraph.

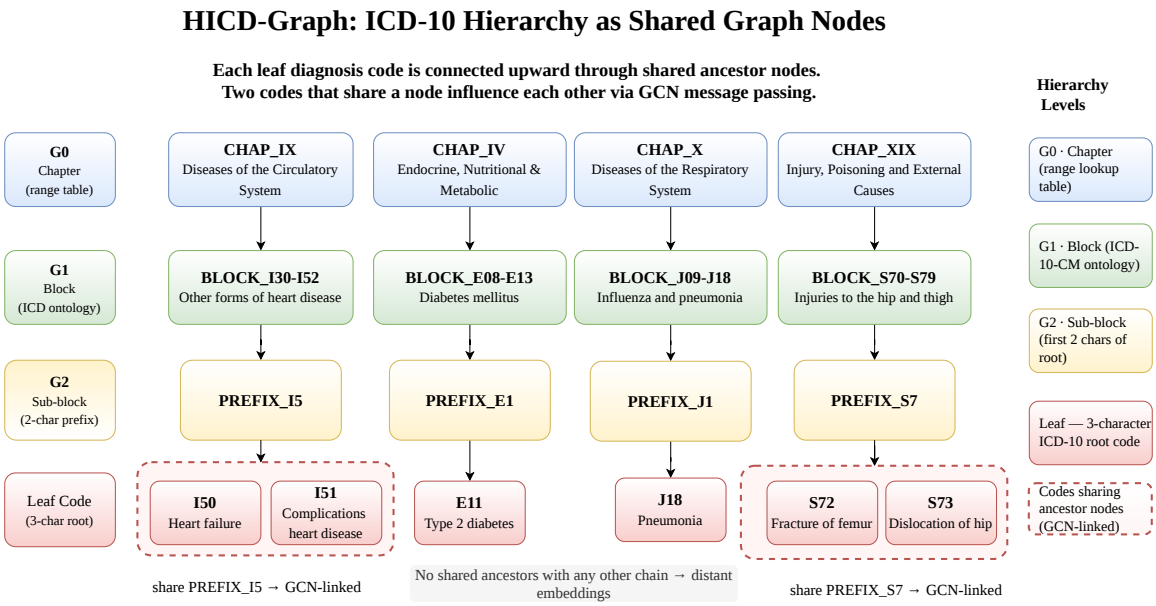


Figure 4: Example of the `HICD-Graph` hierarchy subgraph.

Appendix B. Implementation Details

B.1. `HICD-BERT` Architecture and Training

`HICD-BERT` extends a BEHRT-style encoder by summing diagnosis-token, age, position, and segment embeddings, with optional additive embeddings for the three hierarchy levels `G0`, `G1`, and `G2`. The contextualized sequence is processed by a stack of Transformer blocks, and the pretraining objective is masked language modeling (MLM). For downstream prediction, the encoder is initialized from the pretrained checkpoint and a task-specific multilayer perceptron is applied to the final `[CLS]` representation.

B.1.1. Architecture

`HICD-BERT` builds on a BEHRT-style Transformer encoder and augments the standard token, age, position, and segment embeddings with up to three hierarchy-aware embedding channels corresponding to `G0`, `G1`, and `G2`. These hierarchy embeddings are added to the

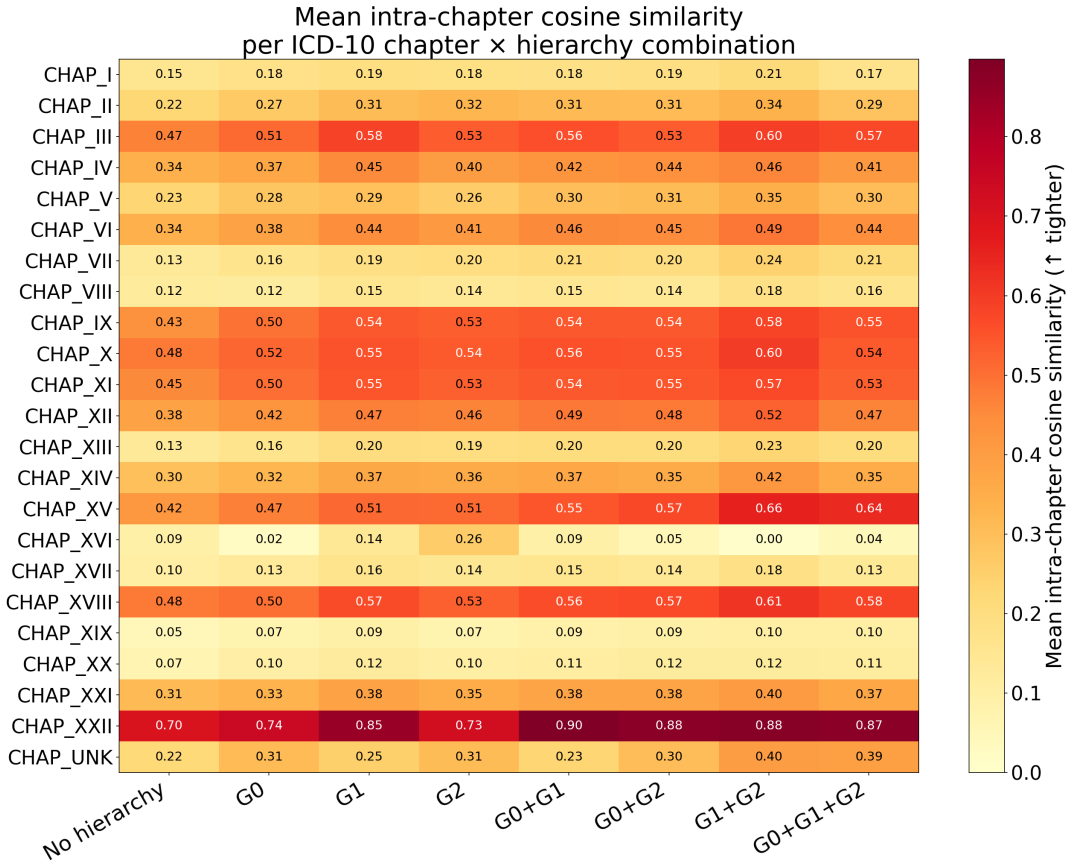


Figure 5: Mean intra-chapter cosine similarity among ICD-10 leaf codes in our pretraining data of MIMIC-IV, computed across all eight hierarchy configurations (`G0`/`G1`/`G2` ablations). Each row corresponds to one ICD-10 chapter, and each column to a hierarchy configuration ordered as in the main results tables. Higher values (darker) indicate that codes within the same chapter are embedded more tightly together. Hierarchy increases similarity across most chapters, with the largest gains in diverse chapters such as Chapter XIX (*Injury and poisoning*, 5,937 codes).

input representation and are learned jointly with the rest of the model. For downstream prediction, the contextualized `[CLS]` representation is passed through a task-specific multilayer perceptron classifier.

B.1.2. MLM Pretraining

We pretrain `HICD-BERT` with a masked language modeling objective over diagnosis sequences. The model is optimized to recover masked diagnosis tokens from their clinical context while jointly conditioning on age, position, segment, and the enabled hierarchy-level embeddings. Table 6 summarizes the pretraining configuration.

Table 5: `HICD-BERT` architecture details.

Component	Setting
Backbone	BEHRT-style Transformer encoder
Embedding components	Token + age + position + segment + optional $G0/G1/G2$ embeddings
Hidden size (d_{model})	512
Transformer layers	6
Attention heads	8
Feed-forward dimension	2048
Dropout	0.1
Maximum sequence length	256
MLM head	Linear $\rightarrow$ GELU $\rightarrow$ LayerNorm $\rightarrow$ Decoder
Task head	3-layer MLP on `[CLS]` representation

Table 6: `HICD-BERT` MLM pretraining hyperparameters.

Hyperparameter	Value
Train batch size	256
Validation batch size	128
Number of workers	8
Epochs	100
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	0.01
Scheduler	CosineAnnealingLR
Scheduler $T_{\max}$	100
Minimum learning rate	1×10^{-6}
Loss	CrossEntropyLoss

B.1.3. Task Fine-Tuning

For downstream tasks, we initialize the encoder from the best MLM checkpoint and optimize a task-specific classification objective. Binary tasks use `BCEWithLogitsLoss` with a positive class weight $w^+ = N_{\text{neg}}^{\text{train}}/N_{\text{pos}}^{\text{train}}$, computed from the training split, to address label imbalance. The best checkpoint is selected by validation AUROC. For binary tasks, a decision threshold is swept over $[0.01, 0.99]$ (200 steps) on the validation set and stored in the checkpoint for test-set evaluation. Table 7 lists the full fine-tuning hyperparameters.

B.2. `HICD-Graph`: Graph-Augmented Hierarchy Model

The graph-based model has two stages. First, we construct a diagnosis co-occurrence graph over ICD roots using PMI-weighted edges, then augment the graph with hierarchy edges corresponding to the enabled `G0`, `G1`, and `G2` levels. Second, we train a two-layer GCN for node embedding learning and use the resulting embeddings to initialize a patient-level Transformer.

Table 7: `HICD-BERT` downstream fine-tuning hyperparameters.

Hyperparameter	Value
Train batch size	128
Validation batch size	256
Number of workers	8
Epochs	64
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	0.01
Scheduler	ReduceLROnPlateau (mode: max)
Scheduler patience	2
Scheduler factor	0.5
Early stopping patience	8
Early stopping min delta	10^{-4}
Early stopping monitor	Validation AUROC
Model selection criterion	Best validation AUROC
Binary-task loss	BCEWithLogitsLoss ($w^+ = N_{\text{neg}}/N_{\text{pos}}$)
Threshold selection	Sweep $[0.01, 0.99]$, 200 steps, max F1 on val set

B.2.1. Graph Construction

We construct an undirected diagnosis graph whose nodes correspond to ICD root codes. Horizontal edges are weighted by PMI computed from within-admission co-occurrence, and vertical edges are added to encode enabled hierarchy levels. Table 8 summarizes the graph construction settings.

Table 8: Diagnosis graph construction hyperparameters.

Hyperparameter	Value
Diagnosis vocabulary	ICD-10-CM codes
ICD-9 handling	Mapped to ICD-10-CM
Horizontal edges	PMI-weighted co-occurrence edges
PMI filtering	Positive PMI only
PMI cutoff	52nd percentile
Hierarchy levels	`G0`, `G1`, `G2` ablated independently
Hierarchy edge weight	1.0
Graph type	Undirected weighted graph

B.2.2. GCN Embedding Training

We learn node embeddings with a two-layer GCN trained using a link-prediction objective over observed and sampled non-edge pairs. The resulting embeddings are used to initialize the downstream patient encoder. The training configuration is given in Table 9.

Table 9: GCN embedding training hyperparameters.

Hyperparameter	Value
Input node feature dimension	64
Input feature initialization	Random normal
GCN layers	2
Hidden dimension	128
Output embedding dimension	128
Edge weights	Min-max normalized to $[0, 1]$
Training objective	Link prediction
Positive pairs	Existing graph edges
Negative pairs	Random non-edge pairs
Loss	BCEWithLogitsLoss
Optimizer	Adam
Learning rate	0.005
Epochs	400

B.2.3. Transformer Fine-Tuning on Graph Embeddings

The learned graph embeddings initialize the diagnosis embedding table of a patient-level Transformer operating over temporally ordered visits. Visit representations are formed by masked mean pooling over code embeddings, and patient representations are obtained through attention pooling across visits. Table 10 provides the fine-tuning details.

B.3. Baseline Implementation Details

We compare against three baselines that share the pretraining, cohort, and evaluation pipeline with our two HICD variants. All baselines are trained on the same patient-level split as the corresponding HICD track, use the same threshold-selection procedure (sweep over $[0.01, 0.99]$ in 200 steps, select the val-set F1-macro maximiser), and are evaluated on the same test set.

No-hierarchy baseline. The no-hierarchy baseline for each architecture is the corresponding HICD model with all hierarchy signals disabled ($(G_0, G_1, G_2) = (0, 0, 0)$). For `HICD-BERT` this removes the additive $G_0/G_1/G_2$ embeddings, leaving a standard BEHRT-style encoder; for `HICD-Graph` this removes the hierarchy edges and ancestor nodes, leaving the PMI-only co-occurrence graph over leaf ICD codes. All other components (tokenizer, pretraining objective, patient-encoder depth/width, fine-tuning schedule) are held fixed. Thus the no-hierarchy baseline is a strict architectural ablation, not a separate model family.

GRAM (Choi et al., 2017). We add GRAM as an ontology-aware recurrent baseline. Each diagnosis code is represented by an attention-weighted aggregation of its ontology ancestors, and the resulting sequence is encoded with a single-layer GRU followed by a linear classifier. We use the ICD-10-CM hierarchy instead of the ICD-9 ontology used in the original paper, with ancestors matched to the corresponding HICD hierarchy levels. ICD-9 codes are mapped through the same GEMs mapping used for HICD. We follow the original GRAM design. Hyperparameters are embedding dimension 128, hidden dimension

Table 10: Patient-level Transformer hyperparameters for the graph-based model.

Hyperparameter	Value
Embedding initialization	GCN node embeddings
Embedding dimension	128
Maximum visits per patient	10
Code aggregation within visit	Masked mean pooling
Visit positional encoding	Learnable
Transformer layers	2
Attention heads	4
Feed-forward dimension	256
Dropout	0.3
Pooling across visits	Attention pooling
Classifier	Linear layer on pooled patient representation
Batch size	64
Epochs	15
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	1×10^{-5}
Scheduler	ReduceLROnPlateau (mode: max)
Scheduler patience	2
Scheduler factor	0.5
Model selection criterion	Best validation AUROC
Loss	BCEWithLogitsLoss ($w^+ = N_{\text{neg}}/N_{\text{pos}}$)
Class balancing	WeightedRandomSampler + w^+
Threshold selection	Sweep $[0.01, 0.99]$, 200 steps, max F1 on val set
Gradient clipping	Max norm 1.0

200, attention dimension 200, dropout 0.5, batch size 100, and maximum sequence length 256.

LR Rollup (Singh et al., 2014). We add the propagated-binary ICD rollup baseline. For each visit, every observed leaf diagnosis activates both its own feature and all ancestor features along the ICD-10-CM hierarchy. These rollup features are used in an ℓ_2-regularized logistic regression classifier with validation-tuned regularization and class weighting.

B.4. Statistical Testing

For all pairwise comparisons of AUROC between a hierarchy-augmented variant and the no-hierarchy baseline, we use the paired DeLong test (DeLong et al., 1988), which provides a closed-form test for the difference between two correlated AUROC estimates. We use reference implementation of Sun and Xu (Sun and Xu, 2014)[1]. We report significance at the $p < 0.05$ threshold.

1. https://github.com/yandexdataschool/roc_comparison

B.5. Hierarchy Ablations

For both `HICD-BERT` and the graph-augmented model, we evaluate all eight hierarchy settings:

$$(0,0,0),\ (0,0,1),\ (0,1,0),\ (0,1,1),\ (1,0,0),\ (1,0,1),\ (1,1,0),\ (1,1,1).$$

Here, $(0,0,0)$ corresponds to the non-hierarchical baseline, while $(1,1,1)$ uses the full hierarchy. In `HICD-BERT` these settings determine which hierarchy embeddings are added to the token representation. In the graph-based model, they determine which hierarchy edges are added to the diagnosis graph and which ancestor nodes are included during sequence tokenization.

B.6. Additional Cross-Dataset Analysis: Relative AUROC Drop

Absolute AUROC values are not directly comparable between MIMIC-IV 30-day inpatient readmission and eICU ICU readmission, because the two settings differ in cohort definition, label semantics, and class prevalence. Because the encoder is held frozen during transfer and only a lightweight task head is fitted on eICU labels, the most informative test of whether hierarchy helps under distribution shift is the within-architecture comparison across hierarchy variants reported in Table 4. As an additional sanity check, we report the relative AUROC drop

$$\Delta_{\text{rel}} = \frac{\text{AUROC}_{\text{MIMIC}} - \text{AUROC}_{\text{eICU}}}{\text{AUROC}_{\text{MIMIC}}},$$

between the in-domain readmission AUROC and the eICU transfer AUROC for every hierarchy variant (Table 11).

Table 11: Per-variant relative AUROC drop between MIMIC-IV 30-day readmission (in-domain) and eICU ICU readmission (frozen-encoder probe). MIMIC and eICU AUROCs are point estimates from the canonical CI-bootstrapped runs.

	`HICD-BERT`			`HICD-Graph`		
Variant	MIMIC	eICU	Δ_{rel}	MIMIC	eICU	Δ_{rel}
No hierarchy	0.6808	0.5653	17.0%	0.6992	0.5645	19.3%
G0	0.6856	0.5711	**16.7%**	0.7086	0.5757	18.8%
G1	0.6810	0.5641	17.2%	0.7079	0.5783	18.3%
G2	0.6920	0.5650	18.4%	0.7086	0.5832	17.7%
G0+G1	0.6862	0.5682	17.2%	0.7105	0.5819	18.1%
G0+G2	0.6917	0.5667	18.1%	0.7097	0.5788	18.5%
G1+G2	0.6916	0.5683	17.8%	0.7113	0.5738	19.3%
G0+G1+G2	0.6921	0.5646	18.4%	0.7114	0.5871	**17.5%**

For both architectures, the best HICD variant on the eICU task achieves a higher absolute eICU AUROC *and* a smaller relative drop than the no-hierarchy baseline (17.5% for `G0+G1+G2` vs. 19.3% for no-hierarchy in `HICD-Graph`). The remaining variants show variable relative drops, in some cases comparable to or larger than the no-hierarchy baseline, reflecting the fact that the underlying MIMIC and eICU prediction settings are not strictly

aligned. We therefore interpret Δ_{rel} as supplementary to the absolute eICU AUROC results in Table 4 rather than a primary measure of transferability; the two analyses agree that hierarchy preserves and in the best variants slightly improves the cross-dataset signal rather than degrading it.

B.7. Per-Chapter ΔAUROC

To further complement the chapter-wise embedding analysis, in this subsection we investigate *where* along the ICD-10 ontology hierarchy yields its largest gains by computing per-chapter ΔAUROC between the best hierarchy variant and the no-hierarchy ablation for each in-domain task and each architecture, restricted to test patients whose visit set is dominated by codes in a given ICD-10 chapter. Tables 12 and 13 report ΔAUROC for the eight highest-volume chapters present across all four tasks; cells marked with * are significant (95% bootstrap CI excludes 0).

Table 12: Per-chapter ΔAUROC for `HICD-BERT` (best hierarchy variant minus no-hierarchy). * denotes 95% bootstrap CI excludes 0.

Chapter	Readmit	EMvisit	LongLOS	Mortality
II Neoplasms	−0.001	+0.024	+0.004	+0.010
IV Endocrine	+0.013*	+0.031*	+0.012*	+0.043*
V Mental	+0.009	+0.041*	+0.023*	+0.003
IX Circulatory	+0.018*	+0.022*	+0.011*	+0.020*
XI Digestive	+0.001	+0.034*	+0.013*	+0.013*
XIII Musculoskeletal	+0.006	+0.011	+0.026	+0.034
XVIII Symptoms	+0.005	+0.050*	+0.010	+0.028*
XXI Health services	+0.011*	+0.027*	−0.001	−0.007

Table 13: Per-chapter ΔAUROC for `HICD-Graph` (best hierarchy variant minus no-hierarchy). * denotes 95% bootstrap CI excludes 0.

Chapter	Readmit	EMvisit	LongLOS	Mortality
II Neoplasms	+0.015	+0.020*	+0.030*	+0.024*
IV Endocrine	+0.017*	+0.008*	+0.008	+0.044*
V Mental	+0.004	+0.005*	+0.007	+0.084*
IX Circulatory	+0.017*	+0.009*	+0.012*	+0.029*
XI Digestive	+0.013*	+0.013*	+0.018*	+0.068*
XIII Musculoskeletal	+0.021	+0.007	+0.048*	+0.065*
XVIII Symptoms	+0.013*	+0.003	+0.011	+0.022*
XXI Health services	+0.012*	+0.007*	+0.016*	+0.068*

Analysis. Positive and relatively consistent improvements occur in high-volume and clinically coherent chapters such as *Circulatory* (Ch. IX), *Endocrine* (Ch. IV), and *Digestive* (Ch. XI), which show positive ΔAUROC across tasks, with significance in many settings,

suggesting that hierarchy may be particularly helpful for predictions involving clinically related diagnosis codes. *Mental* (Ch. V) and *Musculoskeletal* (Ch. XIII) chapters also benefit, though less consistently across tasks-models. Gains are smaller or unstable in heterogeneous chapters such as *Neoplasms* (Ch. II) and chapters dominated by transient or administrative codes (e.g. Ch. XXI). Overall, ICD hierarchy appears most useful in clinically dense chapters where many codes describe related disease processes, potentially supporting information sharing across clinically similar diagnoses, and less useful in more heterogeneous chapters.

B.8. Tokenizer and Hierarchy Mapping

The ICD-code tokenizer and the `G0/G1/G2` ancestor mappings are precomputed once from the MIMIC-IV training cohort and kept fixed throughout MLM pretraining, downstream fine-tuning on each task, and eICU transfer. Any eICU code outside the MIMIC-IV vocabulary is mapped to the `[UNK]` token, and its hierarchy ancestors are looked up from the frozen mapping; no tokenizer or hierarchy retraining occurs at transfer time.

B.9. Cohort Construction, Per-Task Statistics, and ICD-9→10 Mapping Coverage

MIMIC-IV cohort and splits. We use MIMIC-IV v2.2 patients with at least one inpatient stay and an ICD-10 diagnosis (>200K patients). Splits are patient-level 80/10/10, yielding approximately 179K training, 22K validation, and 22K test patients, and no patient appears in more than one split. The same cohort and split are reused for all four MIMIC-IV tasks; per-task differences arise only from label availability. HICD-Graph uses a slightly different visit-eligibility filter, yielding 201K training/val and 22K evaluation patients.

Per-task prediction points and label prevalence. We report test-set metrics over per-visit prediction points, since a single patient can contribute more than one prediction point when the task label is defined at multiple visits in their history. For `HICD-BERT` each task yields ≈32K test prediction points. Positive-class prevalence varies substantially across tasks: emergency admission is majority-positive (≈90%), 30-day readmission is moderately balanced (≈34%), and long length-of-stay and 1-year mortality are strongly imbalanced (≈19% and ≈15% respectively). Because the tasks span mild to severe imbalance, we report AUROC as the primary metric and complement it with F1-macro.

eICU cohort. The eICU transfer cohort consists of ICU unit-stay records derived from the eICU Collaborative Research Database, filtered to patients with at least one diagnosis code and a well-defined ICU-readmission label. eICU diagnosis codes are mapped from ICD-9-CM to ICD-10-CM using CMS GEMs forward mappings (Centers for Medicare and Medicaid Services, 2018) before being looked up against the (frozen) MIMIC-IV vocabulary.

ICD-9→ICD-10 mapping coverage. MIMIC-IV records contain a mix of ICD-9-CM and ICD-10-CM codes; we map all ICD-9 codes to ICD-10-CM via the publicly available General Equivalence Mappings (GEMs). Across the MIMIC-IV cohort, the resulting mapping achieves 95.4% stay-diagnosis coverage, 97.6% stay-level coverage (at least one diagnosis successfully mapped per stay), and 87.5% unique-code coverage. For the eICU

transfer pipeline, after the full ICD-9 → ICD-10 → MIMIC-vocabulary mapping, 95.4% of stay–diagnosis records remain, 97.6% of ICU stays retain at least one diagnosis code, and unique-code coverage is 87.5%. Unmapped codes are assigned to a reserved bucket that is held constant across all hierarchy configurations and therefore cannot systematically explain differences between variants.

B.10. Long Length-of-Stay: Full Results and Analysis

Table 14 reports the full eight-variant ablation for Long-LOS for both `HICD-BERT` and `HICD-Graph`, including the GRAM and LR Rollup baselines.

For `HICD-BERT`, the finest level drives most of the AUROC gain, and combining mid- and fine-level hierarchy is best. The two-level `G1+G2` configuration achieves the strongest AUROC (0.6840), and the all-levels `G0+G1+G2` configuration achieves the strongest F1-macro (0.5787). Among single-level configurations, the finest level `G2` is the only one to significantly improve AUROC over no-hierarchy (0.6808 vs. baseline 0.6728, DeLong $p < 10^{-4}$); the coarser levels `G0`, `G1`, and the two-level `G0+G1` do not significantly move AUROC, even though most variants significantly improve F1-macro. This mirrors the pattern observed in the readmission and emergency-admission tasks: token-level hierarchy injection benefits most from fine-grained groupings that resemble the leaf-code vocabulary, with coarser groupings contributing only when combined with the finest level.

For `HICD-Graph`, every hierarchy variant significantly improves over no-hierarchy. All seven hierarchy configurations significantly improve AUROC over the no-hierarchy baseline, and the best AUROC is achieved by the all-levels `G0+G1+G2` configuration (0.6840), closely followed by `G0+G2` (0.6837). The strongest F1-macro is achieved by `G1+G2` (0.5675). Unlike the strong monotonic pattern observed in readmission and emergency-admission, the gains across single-, two-, and all-level variants are tightly clustered, suggesting that for long-stay prediction the marginal value of adding hierarchy levels saturates once any one level is present.

B.11. 1-Year Post-Discharge Mortality: Full Results and Analysis

Table 15 reports the full eight-variant ablation for 1-year post-discharge mortality for both `HICD-BERT` and `HICD-Graph`, including the GRAM and LR Rollup baselines.

For `HICD-BERT`, the two-level `G0+G2` configuration achieves the best AUROC. The strongest AUROC is achieved by the two-level `G0+G2` configuration (0.8357), while the strongest F1-macro is achieved by the all-levels `G0+G1+G2` configuration (0.7043). Among single-level configurations, the finest level `G2` (0.8305) and the coarsest level `G0` (0.8269) both significantly improve AUROC, whereas the mid-level `G1` alone significantly decreases AUROC relative to no-hierarchy (0.8114 vs. 0.8182). Configurations that combine `G1` with another level recover and exceed the no-hierarchy AUROC, but the strongest AUROC is obtained by combining the coarse and fine levels without `G1`.

For `HICD-Graph`, every hierarchy variant significantly improves over no-hierarchy. All seven hierarchy variants statistically improve AUROC over the no-hierarchy baseline, and the all-levels `G0+G1+G2` configuration achieves both the best AUROC (0.8275) and

Table 14: Long length-of-stay (> 7 days) results for `HICD-BERT`and `HICD-Graph`across hierarchy ablations. Each cell reports the point estimate with 95% bootstrap CI in brackets; **bold** marks the best hierarchy variant per column. * $p < 0.05$ from pairwise DeLong tests on AUROC and paired bootstrap tests on F1-macro, both against the no-hierarchy baseline.

Group	Method	`HICD-BERT`		`HICD-Graph`	
		F1-macro	AUROC	F1-macro	AUROC
Baselines	No hierarchy	0.5337 [.528,.539]	0.6728 [.665,.680]	0.5374 [.532,.544]	0.6712 [.664,.679]
	GRAM	0.5475 [.542,.553]	0.6752 [.669,.682]	0.5584 [.552,.565]	0.6729 [.665,.681]
	LR rollup	0.5337 [.528,.539]	0.6298 [.623,.637]	0.5475 [.542,.553]	0.6381 [.630,.646]
Single-level	G0	0.5667* [.561,.573]	0.6733 [.666,.681]	0.5511* [.545,.557]	0.6815* [.674,.689]
	G1	0.5692* [.564,.575]	0.6718 [.664,.679]	0.5615* [.556,.567]	0.6830* [.676,.691]
	G2	0.5720* [.567,.578]	0.6808* [.673,.688]	0.5586* [.553,.564]	0.6825* [.675,.690]
Two-Level	G0+G1	0.5647* [.559,.570]	0.6735 [.666,.681]	0.5615* [.556,.567]	0.6830* [.676,.691]
	G0+G2	0.5600* [.554,.565]	0.6809* [.674,.688]	0.5406 [.535,.547]	0.6837* [.676,.691]
	G1+G2	0.5578* [.552,.563]	**0.6840*** [.677,.691]	**0.5675*** [.562,.573]	0.6819* [.675,.689]
All-Levels	G0+G1+G2	**0.5787*** [.573,.584]	0.6814* [.674,.689]	0.5538* [.548,.559]	**0.6840*** [.677,.691]

the best F1-macro (0.6961). The absolute gain over no-hierarchy (+0.0393 AUROC) is the largest improvement observed across all four in-domain tasks for either architecture, and the progressive gain from single-level to two-level to all-levels variants is pronounced, mirroring the monotonic pattern seen in readmission.

Across both Long-LOS and Mortality, both `HICD-BERT` and `HICD-Graph` match or outperform the GRAM and LR Rollup baselines on AUROC. The mortality task shows the clearest distinction between the two architectures: `HICD-BERT` achieves its best AUROC using the selective coarse-and-fine `G0+G2` combination, whereas `HICD-Graph` achieves its best AUROC and F1-macro by integrating all hierarchy levels.

Table 15: 1-year post-discharge mortality results for `HICD-BERT`and `HICD-Graph`across hierarchy ablations. Each cell reports the point estimate with 95% bootstrap CI in brackets; **bold** marks the best hierarchy variant per column.* $p < 0.05$ from pairwise DeLong tests on AUROC and paired bootstrap tests on F1-macro, both against the no-hierarchy baseline.

Group	Method	`HICD-BERT`		`HICD-Graph`	
		F1-macro	AUROC	F1-macro	AUROC
Baselines	No hierarchy	0.6861 [.679,.692]	0.8182 [.812,.824]	0.6765 [.670,.683]	0.7882 [.782,.795]
	GRAM	0.6970 [.691,.704]	0.8270 [.821,.833]	0.6516 [.645,.658]	0.7567 [.749,.764]
	LR rollup	0.6603 [.654,.666]	0.7685 [.762,.776]	0.6831 [.677,.689]	0.7958 [.789,.803]
Single-level	G0	0.6989* [.692,.706]	0.8269* [.821,.833]	0.6848* [.679,.691]	0.8174* [.811,.824]
	G1	0.6902 [.683,.697]	0.8114* [.805,.818]	0.6825* [.676,.689]	0.8176* [.811,.824]
	G2	0.6970* [.691,.703]	0.8305* [.825,.837]	0.6859* [.679,.692]	0.8219* [.816,.828]
Two-Level	G0+G1	0.6971* [.690,.703]	0.8276* [.822,.834]	0.6904* [.684,.697]	0.8249* [.818,.831]
	G0+G2	0.7025* [.696,.708]	**0.8357*** [.830,.842]	0.6937* [.687,.700]	0.8265* [.820,.833]
	G1+G2	0.7036* [.697,.710]	0.8313* [.826,.837]	0.6865* [.680,.693]	0.8196* [.813,.826]
All-Levels	G0+G1+G2	**0.7043*** [.698,.711]	0.8333* [.827,.839]	**0.6961*** [.690,.703]	**0.8275*** [.821,.834]

B.12. Reproducibility Statement

To support reproducibility we provide a complete specification of both HICD variants and preprocessing in this appendix, and our code is available here. In particular, we detail the `HICD-BERT` architecture and MLM/fine-tuning hyperparameters in Appendix B.1, the `HICD-Graph` graph construction, GCN embedding, and patient-encoder hyperparameters in Appendix B.2, and the GRAM, LR Rollup, and no-hierarchy baseline configurations in Appendix B.3. The preprocessing pipeline, ICD-9→ICD-10 mapping, cohort construction, and per-task label prevalence are documented in Appendix B.9, and the paired DeLong / paired bootstrap evaluation harness is described in Appendix B.4. The datasets, MIMIC-IV and eICU are available to credentialed researchers via PhysioNet under the standard data-use agreement.