SMI: Semantic Medical ID for Hierarchy-Aware Concept Representation

Lia Shen[1] **Qincheng Lu**[2] **He Zhu** [2] **Ziyang Song**[3] *

[1] School of Computer Science, University of Wisconsin–Madison [2] School of Computer Science, McGill University
[3] School of Electrical Engineering and Computer Science (EECS), Ohio University

Abstract

Recent advances in generative AI have accelerated the use of language models (LMs) for clinical prediction tasks. However, existing biomedical LMs often struggle to capture clinically meaningful relationships among medical concepts, as they rely solely on data-driven text learning and overlook domain knowledge. In this study, we propose **Semantic Medical ID (SMI)**, a novel representation framework that integrates an expert-defined medical ontology into LM-based embeddings. By leveraging the hierarchical structure of medical ontologies, SMIs generate embeddings that preserve clinical relationships across major disease categories, subcategories, and specific conditions, enhancing interpretability for clinical end users. Experimental results demonstrate that SMI improves predictive accuracy in mortality and readmission tasks. SMI also exhibits greater robustness under cross-hospital distribution shifts, highlighting its effectiveness in producing clinically generalizable representations.

1 Introduction

The widespread adoption of electronic health record (EHR) systems has generated large-scale clinical data, enabling machine learning (ML) models to support tasks such as diagnosis prediction, drug recommendation, and patient risk stratification. Recent advances in language models (LMs) have shown promise in processing and interpreting clinical text. To adapt LMs for healthcare, they are either fine-tuned on clinical data after being pre-trained on general corpora Singhal et al. [2023, 2025], or directly pre-trained on biomedical datasets Lee et al. [2020]. However, current biomedical LMs struggle to capture clinically meaningful representations of structured medical concepts. For instance, the disease conditions are encoded using the International Classification of Diseases (ICD) coding system Steindel [2010]. However, directly encoding the ICD codes based on their text descriptions fails to capture the underlying clinical relationships between clinical codes.

Despite the hierarchical organization of ICD codes (Fig. 2.a), current biomedical LMs struggle to capture these structural relationships. In Fig. 1, we evaluate whether the embeddings of child concepts are closer to their parent concepts than to their non-parent ones, assessing the model's ability to capture clinical hierarchy. Using expert-defined parent-child pairs across three levels, BioBERT achieves 85.1% (Major $\rightarrow$ Sub-category), 74.7% (Sub-category $\rightarrow$ CCS), but only 66.0% (CCS $\rightarrow$ ICD), respectively. While the model captures coarse-grained relationships, its ability to distinguish fine-grained medical concepts drops significantly. Biomedical LMs often struggle to understand fine-grained medical concepts due to limited training data for rare conditions and the absence of structured domain knowledge Song et al. [2025]. This limits their ability to model hierarchical semantics, reducing their effectiveness in tasks requiring nuanced medical understanding. Clinical

*Correspondence to ziyangs@ohio.edu.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFM.

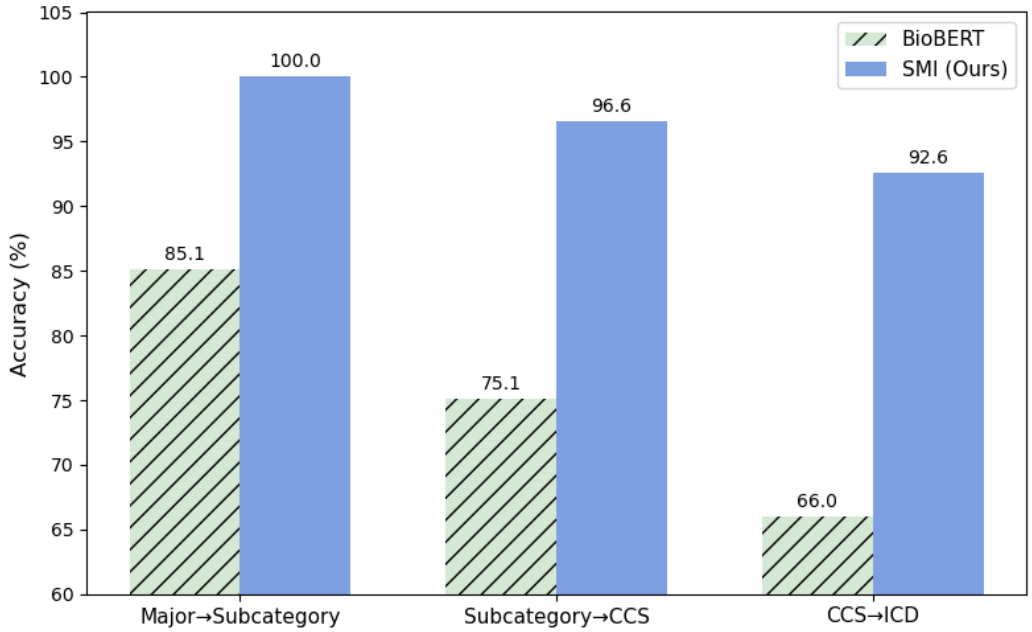

Figure 1: **Comparison of Hierarchical Similarity Accuracy.** We evaluate whether the embedding of a child concept is more similar to its parent than to non-parent concepts across three hierarchical levels: Major→Sub-category, Sub-category→CCS, and CCS→ICD. Our proposed SMI embedding substantially improves discriminative accuracy, especially at finer granularity (e.g., from 66.0% to 92.6% on CCS→ICD level), demonstrating stronger understanding of hierarchical medical semantics.

coding systems such as ICD organize medical concepts hierarchically, from broad disease categories to fine-grained conditions. Integrating such domain knowledge can help LMs better capture the hierarchical semantics among medical concepts.

Recently, recommendation systems have adopted semantic ID framework to generate hierarchy-aware and semantically meaningful item representations Ju et al. [2025], Singh et al. [2024], Rajput et al. [2023]. It uses RQ-VAE to hierarchically assign discrete codes that capture multi-level semantics. Although the semantic ID technique can generate hierarchy-aware representation Lee et al. [2022], it requires to predefined hierarchy depth and equal token counts per level. Moreover, it relies on data-driven learning and lacks integration of expert-defined hierarchies, resulting in less interpretable and meaningful hierarchical semantic representations. In this study, we propose **Semantic Medical ID (SMI)** that integrates expert-defined hierarchical structures from medical ontology to generate semantic embeddings. Using a four-level hierarchy from ICD coding system, SMI constructs semantic IDs and hierarchy-aware embedding for medical concepts. We quantitatively evaluate SMI and show that it offers stronger interpretability by explicitly encoding hierarchical semantics from domain knowledge. Visualization results reveal that SMI learns hierarchy-aware embeddings that organize medical concepts aligned with domain knowledge and forms distinct clusters. On the MIMIC-III dataset, SMI outperforms biomedical LMs in multiple clinical prediction tasks. Cross-hospital evaluation on the eICU dataset further confirms its robustness to distribution shifts, achieving better generalization across sites.

2 Background

2.1 Medical Ontology

Medical ontologies, such as ICD, are developed by domain experts to organize clinical concepts into hierarchical structures that capture semantic relationships from broad disease categories to fine-grained conditions Steindel [2010]. However, most ML models treat each medical code as an independent token, ignoring this rich structural information. For instance, ML models often

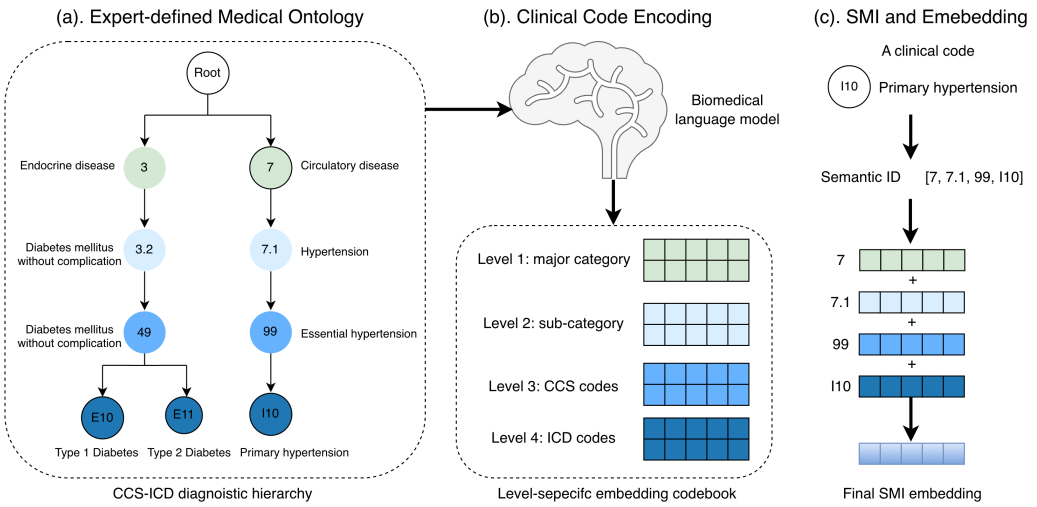

Figure 2: Outline of the SMI embedding process. **a**. Expert-defined medical ontology organizes clinical concepts from coarse to fine granularity. **b**. A biomedical language model encodes clinical codes at each level into level-specific embedding spaces. **c**. For each ICD code, its SMI is derived from the expert-defined ontology. Its embedding is aggregated across various levels to form the final representation, capturing its hierarchical semantics.

struggle with long-tail concepts (e.g., rare diseases) that share semantics with more prevalent ones. To overcome this limitation, we propose a Semantic ID framework that explicitly incorporates an expert-defined medical ontology into a structured, multi-level representation, enabling the model to capture clinically meaningful relationships and improve generalization across diseases.

2.2 Semantic ID Methodology

In recommendation systems, semantic ID provides a structured representation of discrete tokens, capturing coarse-to-fine semantic hierarchies among item IDs Ju et al. [2025], Singh et al. [2024], Rajput et al. [2023]. Prior methods employ RQ-VAE to construct these representations through vector quantization with a fixed number of levels and equal tokens per level defined as hyperparameters. It also requires using token parameterization techniques such as prefix-n-grams or modulo hashing. However, such data-driven approaches lack domain knowledge and often produce uninterpretable or clinically irrelevant hierarchy in the medical domain. To address this, we construct Semantic IDs for medical concepts based on an expert-defined medical ontology. By replacing data-driven learning with structured medical knowledge, our method ensures interpretability and improves generalization for downstream clinical tasks.

3 Methodology

In this section, we leverage domain knowledge to build the medical ontology for ICD diagnostic codes. We introduce a two-stage method for learning semantic ID and hierarchy-aware embeddings for ICD codes. We first apply a biomedical LM to encode each concept at four levels in the hierarchy. We then represent each clinical code with a semantic ID, which is a sequence of coarse-to-fine discrete clinical concepts. The resulting embeddings from SMIs will both capture clinical meaning via biomedical LMs and encode their coarse-to-fine clinical semantics.

3.1 Hierarchy from Medical Ontology

Accurately representing medical concepts requires embeddings that reflect their inherent hierarchical semantics. Instead of relying on a data-driven hierarchy learning model such as RQ-VAE, we explicitly incorporate domain knowledge to represent each medical concept (e.g., an ICD code) from coarse to fine granularity. This forms a hierarchical path within a tree-structured clinical

Table 1: Hierarchical mapping of the four-level CCS–ICD taxonomy.

Level	Code index	Code ID	Description
Major	0	1	Infectious and parasitic diseases
:	:	:	:
Sub	0	1.1	Infectious and parasitic diseases → Bacterial infection
:	:	:	:
CCS	0	1	Infectious and parasitic diseases → Bacterial infection → Tuberculosis
:	:	:	:
ICD	0	001	Diseases of the digestive system → Intestinal infection → Intestinal infection → Cholera
:	:	:	:

Table 2: Hierarchy-aware prompt for biomedical LMs. We illustrate an example of ICD-10 diagnostic code **I10 Essential (Primary) Hypertension**, where the text from all hierarchy levels (i.e., Major Category, Sub-category, CCS code, and ICD code) is explicitly concatenated for the biomedical encoder.

Level	Major Category	Sub-category	CCS Code	ICD Code
Prompt	Diseases of circulatory system	Diseases of circulatory system → Hypertension	Diseases of circulatory system → Hypertension → Essential Hypertension	Diseases of circulatory system → Hypertension → Essential Hypertension → Essential (Primary) hypertension

ontology (e.g., Circulatory Disease → Hypertension → Essential Hypertension). To achieve this, we construct a comprehensive hierarchy that unifies both **ICD-9-CM** and **ICD-10-CM** diagnostic systems within a multi-level, expert-defined taxonomy (Fig. 2.a). This clinical ontology defines the structural relationship for generating Semantic Medical IDs (SMIs), allowing embeddings to encode disease relationships consistent with expert-defined medical knowledge.

We first integrate ICD-9-CM and ICD-10-CM into a unified hierarchy to ensure comprehensive coverage of diagnostic concepts across coding systems. Only integer-form codes were retained to simplify hierarchical mapping and ensure unambiguous alignment across ICD versions. The generated unified hierarchy includes 937 ICD-9 codes and 1,278 ICD-10 codes.

To enable clinically interpretable disease categorization, we adopt the **Clinical Classifications Software (CCS)** developed by the Agency for Healthcare Research and Quality (AHRQ). CCS provides an expert-curated framework that groups granular diagnostic codes into broader disease categories, supporting both coarse- and fine-grained clinical analyses. In total, CCS comprises 18 major disease categories, 134 subcategories, and 265 CCS codes, covering all ICD-9-CM and ICD-10-CM diagnostic concepts. Based on this structure, we construct a four-level hierarchy: **Major category** → **Sub-category** → **CCS code** → **ICD code**. As each child is only linked to a single parent in this tree, every ICD code is uniquely assigned to a hierarchical path tracing its lineage across levels (e.g., Diseases of the circulatory system → Hypertension → Hypertension with complications and secondary hypertension → Essential hypertension). This one-to-one mapping ensures that each diagnostic code has a distinct and interpretable position in the taxonomy. The resulting ontology provides a structured foundation for embedding learning, enabling SMIs to reflect expert-defined clinical relationships among diseases.

3.2 Clinical Code Encoding

To obtain clinically meaningful representations, we encode the textual descriptions of clinical codes at four levels of a clinical hierarchy, i.e., major category, sub-category, CCS code, and ICD code (Fig. 2.b). Specifically, we use a frozen BioBERT model as an encoder. For each medical concept, we first generate the description of the clinical code at every level and explicitly concatenate its full hierarchical path, incorporating the descriptions of all ancestor codes in the hierarchy. We find that the hierarchy-aware input is more effective in capturing medical semantics. Table 2 presents the

BioBERT input prompt, in which the text passed to the encoder explicitly encodes the hierarchical context.

Let the hierarchy levels be $l \in \mathcal{L} = \{1, 2, 3, 4\}$, corresponding to a specific level (e.g., Major category, Sub-category, CCS, ICD) with $K^{(l)}$ clinical codes. Our hierarchy has $(K^{(1)}, K^{(2)}, K^{(3)}, K^{(4)}) = (18, 134, 265, 2215)$, yielding a total of 2632 clinical codes. Therefore, the codebook of l-th level is $c^{(l)} = \{ c_i^{(l)} \}_{i=1}^{K^{(l)}}$. With embedding size $d = 768$, we define an embedding table for each level $\mathbf{E}^{(l)} \in \mathbb{R}^{K^{(l)} \times d}$. For each code $c_i^{(l)}$, BioBERT encodes its tokenized input and produces contextualized token embeddings $\text{Enc}(c_i^{(l)})$. We then apply mean pooling across tokens to obtain a single embedding $\mathbf{e}_i^{(l)} \in \mathbb{R}^d$, which occupies row i of $\mathbf{E}^{(l)}$.

3.3 Semantic IDs Generation and Embedding Aggregation

Following prior work, a Semantic ID is a sequence of discrete codes $\mathbf{c} = (c^1, \ldots, c^L)$ produced by the encoder and the expert-defined hierarchy, ordered from coarse to fine granularity. As shown in Fig. 2.c, in our hierarchy, a raw ICD code is mapped to a sequence $(c^{(1)}, c^{(2)}, c^{(3)}, c^{(4)})$. Specifically, the first token $c^{(1)}$ denotes the coarsest concept of major disease category, while $c^{(2)}$ and $c^{(3)}$ specify sub-category and CCS code, respectively. The last token $c^{(4)}$ further refines the most fine-grained ICD code. This allows us to control both the amount and the structure of clinical information encoded within each SMI.

In contrast, prior semantic ID methods based on RQ-VAE construct latent hierarchies through vector quantization, assuming uniform codebook sizes across all levels. However, such data-driven hierarchies often fail to align with established medical ontologies, resulting in poor interpretability and clinically inconsistent groupings. For instance, RQ-VAE may cluster unrelated ICD codes together due to statistical similarities rather than shared medical meaning. Our method explicitly encodes the expert-defined medical hierarchy into a multi-level discrete representation. This ensures that the learned semantic structure reflects true clinical relationships and improves interpretability in downstream healthcare tasks.

Previous approaches using RQ-VAE rely on unsupervised clustering and token parameterization (e.g., prefix n-grams or modulo hashing) to generate latent semantic codes. In contrast, our approach leverages an expert-defined medical hierarchy that maps each ICD code to a fixed sequence of discrete codes. This eliminates the need for token parameterization techniques, as the semantic meaning of each ICD code is explicitly defined. To compute the embedding for each ICD code, we aggregate the embeddings of the code and all its ancestor nodes along the hierarchy path $\mathcal{P}(i)$ using sum pooling:

$$\mathbf{e}_i^{\text{SMI}} = \sum_{(l,j) \in \mathcal{P}(i)} \mathbf{e}_j^{(l)} \tag{1}$$

where $\mathcal{P}(i)$ denotes the set of (l, j) pairs representing the hierarchical lineage of code i across levels $l \in \mathcal{L}$. This results in a single d-dimensional embedding $\mathbf{e}_i^{\text{SMI}}$ for each ICD code, capturing both the fine-grained meaning of the diagnosis and its broader clinical context.

4 Experiments and Results

4.1 Dataset and Preprocessing

MIMIC-III is a publicly available critical care database containing de-identified health records of over 40,000 patients admitted to the intensive care units at Beth Israel Deaconess Medical Center between 2001 and 2012 Johnson et al. [2016]. In this study, we use MIMIC-III for clinical prediction tasks, including mortality and readmission prediction. We extract diagnostic records from 7,537 patients with multiple hospital admissions. Each diagnosis is represented by its ICD-9 code, which is mapped to its integer-level code. For each patient, all diagnoses within a visit are treated as an unordered set, while visits are chronologically ordered.

The eICU Collaborative Research Database is a multi-center intensive care unit (ICU) database containing over 200,000 admissions from ICUs monitored by eICU programs in the United States Pollard et al. [2018]. It offers de-identified EHR data, encompassing demographics, diagnoses, treatments, and interventions. We use the eICU dataset to evaluate distribution shifts in patient

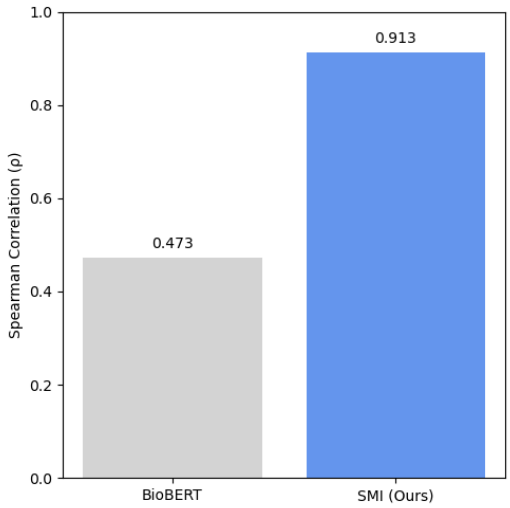

Figure 3: Spearman Correlation Between Embedding Similarity and Hierarchical Proximity. We assess how well SMI and BioBERT embedding methods preserve the expert-defined medical hierarchy by computing the Spearman correlation between embeddings' cosine similarity and hierarchical proximity. Our SMI embeddings achieve a significantly higher correlation (ρ= 0.913) than BioBERT (ρ = 0.473), indicating stronger alignment with the expert-defined medical ontology.

observations across hospitals using both SMI and BioBERT embedding methods. Diagnosis records are mapped to 72 unique integer-level ICD-9 codes. We treat each patient's medical records as unordered observations without considering temporal order. We select the nine hospitals with the largest number of patients, resulting in 2,134 patients per hospital and a total of 19,208 patients for analysis.

4.2 SMI Embeddings Capture Hierarchical Semantics

To evaluate the effectiveness of SMI in capturing hierarchical structure, we assess its ability to preserve parent–child relationships among clinical concepts. Specifically, we evaluate three parent–child hierarchy levels: Major $\rightarrow$ Sub-category, Sub-category$\rightarrow$CCS, and CCS $\rightarrow$ ICD. For each hierarchy level, we iterate over all parent-child pairs and report the average accuracy across ten independent runs. As shown in Fig. 1, SMI consistently outperforms BioBERT across all levels, with substantial improvements at finer granularity. While both models find it more challenging to discriminate fine-grained relationships, SMI successfully distinguishes 92.6% of CCS-ICD pairs compared to 66.0% for BioBERT. This demonstrates that integrating expert-defined medical ontologies allows SMI to produce more interpretable and hierarchically consistent embeddings than biomedical LMs.

We assess whether SMI embeddings preserve hierarchical semantics by measuring the Spearman correlation between ICD embedding similarity and their Least Common Ancestor (LCA) height in the expert-defined hierarchy. The LCA height indicates how closely two medical concepts are related. For instance, "Type 1 diabetes" and "Type 2 diabetes" share the same CCS category, giving them an LCA height of 3. Therefore, the cosine similarity between these two ICD embeddings should be high. A higher LCA height indicates that two ICD codes are more semantically related, and thus their embeddings should exhibit higher similarity. We compute cosine similarities between ICD embeddings and correlate them with LCA heights. We randomly sample 25,000 ICD code pairs for each LCA height (i.e., 1 to 4)and compute the average Spearman correlation over ten runs. As shown in Fig. 3, SMI embeddings achieve a much higher Spearman correlation (ρ=0.913) than BioBERT embeddings (ρ=0.473). This demonstrates that SMI captures hierarchical proximity and expert-defined ontology structures far more effectively. In contrast, BioBERT fails to distinguish embeddings of clinically distant concepts, reflecting its limited understanding of hierarchical medical semantics.

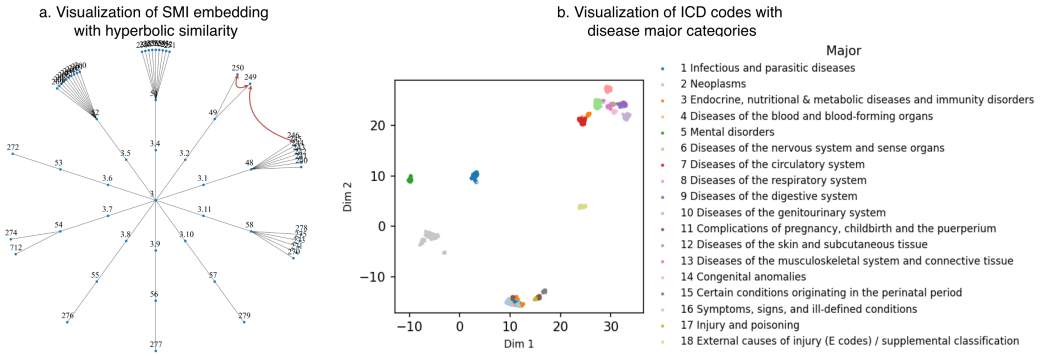

Figure 4: Visualization of the learned ICD diagnostic embeddings. (a) SMI embeddings of all descendant concepts under the Major Disease Category 3 (Endocrine, Nutritional, and Metabolic Diseases and Immunity Disorders). Pairwise relationships are evaluated using hyperbolic similarity, which effectively captures hierarchical distances. The SMI-learned layout preserves the expert-defined hierarchy tree, with child concepts positioned near their parents. (b) UMAP projection of learned ICD embeddings across all 18 major disease categories. The ICD codes form distinct clusters within each major category, aligning closely with expert-defined medical ontology. This demonstrates that the SMI approach effectively captures clinically meaningful and hierarchical semantics encoded in the medical ontology.

4.3 Qualitative Visualization of SMI Embeddings

To qualitatively assess whether SMI preserves the expert-defined medical hierarchy, we visualize the embeddings of all descendant concepts under Major Disease Category 3 (Endocrine, Nutritional, and Metabolic Diseases and Immunity Disorders). As shown in Fig. 4.a., we compute the hyperbolic similarity between each parent–child pair to represent hierarchical relationships among medical concepts. Because the hyperbolic metric expands exponentially with increasing Euclidean distance, it provides a natural geometry for capturing hierarchical structures. For analysis only, we project the learned embedding onto a hyperbolic manifold using the Poincaré ball Nickel and Kiela [2017], defined as $\mathbb{B}^d = \{x \in \mathbb{R}^d \mid \|x\| < 1\}$ with curvature $c = 1.0$. The hyperbolic distance between two embeddings $u, v \in \mathbb{B}^d$ is computed as:

$$d_{\mathbb{B}}(u, v) = \operatorname{arcosh}\left(1 + \frac{2c\|u - v\|^2}{(1 - c\|u\|^2)(1 - c\|v\|^2)}\right) \tag{2}$$

To facilitate visualization and interpretability, we transform the hyperbolic distance into a similarity score, $\mathrm{sim}(u, v) = \exp\left(-\frac{d_{\mathbb{B}}(u,v)}{\tau}\right)$, with $\tau = 1$ is a temperature parameter.

For Major Disease Category 3 (Endocrine, Nutritional, and Metabolic Diseases and Immunity Disorders), we visualize the layout by positioning parent–child pairs according to their hyperbolic similarities, which capture semantic relatedness in the learned embedding space (Fig. 4.a). The SMI embeddings, evaluated through hyperbolic similarity, reveal a hierarchical structure that aligns closely with the expert-defined medical hierarchy. This alignment suggests that SMI leverages medical ontologies to learn hierarchy-aware embeddings that are consistent with domain knowledge. By summing embeddings across hierarchical levels, SMI aggregates multi-level semantic information from the medical ontology, enabling the model to capture structured clinical relationships in the learned embedding space.

We apply Uniform Manifold Approximation and Projection (UMAP) to project the learned embeddings into 2D space for visualization McInnes et al. [2018], Healy and McInnes [2024]. Each point represents an ICD diagnostic code, which is colored by its corresponding major disease category (Fig. 4.b). The visualization reveals distinct clusters, where ICD codes from the same major disease category are grouped closely together. This clear separation across categories indicates that the learned embeddings effectively capture semantic distinctions between disease conditions.

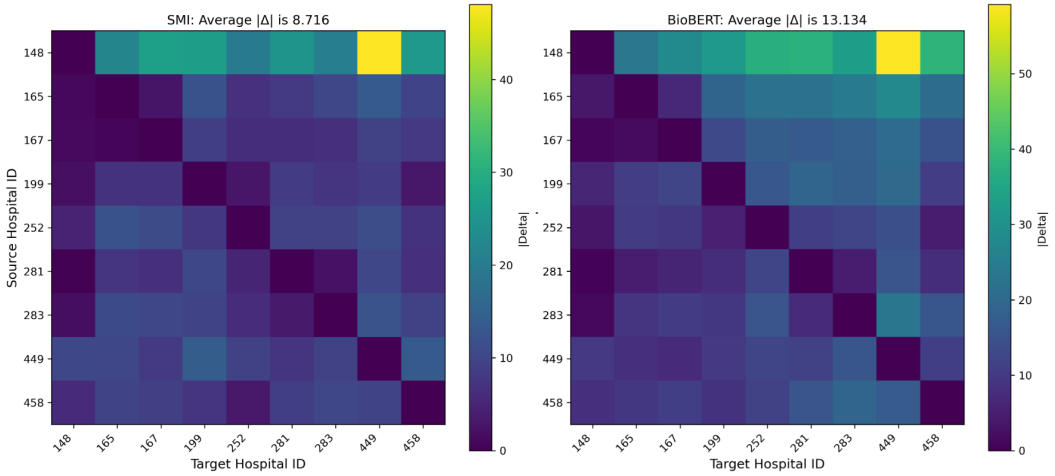

Figure 5: Cross-hospital distribution shift analysis based on average absolute log-likelihood differences ($|\Delta|$) between source and target hospitals. Each heatmap illustrates the distribution shift of patient embedding using SMI (left) and BioBERT (right), both evaluated with a VAE-based density model. SMI embeddings exhibit smaller distribution shifts (average $|\Delta| = 8.7$) compared to BioBERT ($|\Delta| = 13.1$), suggesting better cross-site consistency and generalization.

Table 3: **Clinical Prediction Performance.** Comparison of AUROC and AUPRC scores between SMI and BioBERT embeddings on mortality and readmission prediction tasks. SMI outperforms BioBERT in binary prediction with the aid of hierarchical semantics.

Model	**Mortality**		**Readmission**	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**
SMI	65.32	31.58	62.60	38.65
BioBERT	60.38	28.33	61.45	38.98

4.4 SMI Improves Prediction Performance

We evaluated the utility of learned patient embeddings across multiple admissions on two binary clinical prediction tasks: (1) in-hospital mortality prediction, and (2) 15-day readmission prediction. Each task is formulated as a mapping from the embedding of all previous visits $x_{1:t-1}$ to the patient outcome of the next visit $y[x_t]$:

$$f : (x_{1:t-1}) \rightarrow y[x_t]$$

To construct patient embeddings, we directly use the SMI embeddings of all ICD codes and apply mean pooling across tokens to obtain a single patient-level representation. In contrast, the BioBERT baseline encodes the textual descriptions of all ICD codes as a sequence, using the [CLS] token from the final layer as the patient-level embedding.

We make classification using a linear probing approach with a single linear layer. For binary prediction of mortality and readmission, a linear classifier with a weight $w \in R^{d\times 1}$ and a sigmoid activation is trained using binary cross-entropy loss. All models are implemented in PyTorch and optimized with Adam optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-5}). Training uses a batch size of 64 for up to 20 epochs with early stopping. An 80/20 train–test split is applied, and performance is evaluated using AUROC and AUPRC.

As shown in Table 3, SMI achieves higher predictive performance than BioBERT across both mortality and readmission prediction tasks. Specifically, SMI improves AUROC and AUPRC in mortality prediction (65.32% and 31.58%) compared to BioBERT (60.38% and 28.33%). Notably, our SMI approach requires no additional encoding or fine-tuning, but it simply applies mean pooling of precomputed semantic embeddings. This demonstrates that the hierarchical semantics embedded in SMI enhance clinical representation, leading to more discriminative modeling. These findings

highlight an effective yet simple approach to leveraging domain knowledge for improving clinical outcome prediction, without relying on computationally intensive language models or fine-tuning.

4.5 SMI Mitigates Distribution Shift

To assess cross-hospital distribution shifts, we train a Variational Autoencoder (VAE) as a density estimator on patient embeddings from the eICU dataset Kingma and Welling [2013]. For each source hospital k, we encode patient diagnostic records using either SMI or BioBERT, and then train a VAE to model the data distribution $p_k(x_k)$, where $p_k(\cdot)$ denotes the hospital-specific density estimator learned from the data x_k. We then evaluate generalization to a target hospital t by computing the difference in log-likelihoods:

$$\Delta = \mathbb{E}_{x_k}\left[\log p_k(x_k)\right] - \mathbb{E}_{x_t}\left[\log p_k(x_t)\right] \tag{3}$$

where $\log p_k(x_k)$ and $p_k(x_t)$ denote the log-likelihoods of in-domain samples x_k and out-of-domain samples x_t evaluated using the VAE model trained on k-th hospital, respectively. A large Δ implies greater shift between source and target distributions, indicating cross-hospital generalization. We estimate the log-likelihood $\log p_k(x)$ using the Importance Weighted Autoencoder (IWAE) objective, a tighter bound on the true log-likelihood than the standard ELBO Burda et al. [2015]. For each sample x_i, we draw $K = 64$ importance samples $z_i \sim q(z \mid x)$ to compute:

$$\log p_k(x) \approx \log\left(\frac{1}{K}\sum_{i=1}^{K}\frac{p(x, z^{(i)})}{q(z^{(i)} \mid x)}\right) \tag{4}$$

Fig. 5 illustrates cross-hospital distribution shifts between source and target hospitals. Each heatmap visualizes how well a density model trained on one hospital generalizes to another, where each element represents the absolute difference in log-likelihood ($|\Delta|$) between the source and target hospitals. SMI embeddings produce consistently lower $|\Delta|$ values across hospital pairs compared with BioBERT embeddings. This indicates that SMI better preserves the underlying population structure and reduces domain shift. This improvement can be attributed to the incorporation of expert-defined medical hierarchies into SMI, which embed semantically related diseases in close proximity. By contrast, BioBERT embeddings rely solely on textual correlation and contextual similarity without reflecting clinical ontology relationships, resulting in higher variability across sites. These results demonstrate that SMI effectively mitigates cross-hospital distribution shifts, demonstrating its potential for multi-institutional modeling where data heterogeneity often limits generalization of ML models.

5 Conclusion

In this work, we introduce SMI framework that leverages expert-defined medical ontology to encode clinically meaningful hierarchical semantics into medical concept representations. Specifically, we use a tree-based hierarchy extracted from the medical ontology (e.g., a four-level CCS system) to construct a SMI for each concept, representing it as a sequence along its hierarchical path. By doing so, it represents each medical concept by summing LM-encoded embeddings across multiple hierarchy levels, thereby integrating domain knowledge from the medical ontology into the final representations. SMI provides interpretable and effective representations that improve the prediction of clinical outcome in the MIMIC-III dataset, outperforming biomedical LMs. Moreover, it demonstrates strong generalization across hospitals on the eICU dataset, showing robustness to cross-site distribution shifts.

Currently, our study models a four-level ICD hierarchy using integer-level ICD codes. We plan to extend this to fine-grained ICD codes with decimals. We will also integrate ICD procedure and medication ATC hierarchies to enable analyses of disease–treatment interactions. In addition, the current method learns embeddings in Euclidean space. In future work, we aim to learn embeddings directly in hyperbolic space, which is well suited for representing hierarchical semantics. For experiments, we will expand evaluation to a broader suite of clinical prediction tasks on MIMIC-III and eICU datasets. For distribution-shift analysis, we will move beyond likelihood-based evaluations and assess model robustness through predictive tasks under cross-site settings.

References

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H Chen, Nigam H Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera Y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S Sara Mahdavi, Joelle K Barral, Dale R Webster, Greg S Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nat. Med.*, 31(3):943–950, March 2025.

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

Steven J Steindel. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J. Am. Med. Inform. Assoc.*, 17(3):274–282, May 2010.

Ziyang Song, Qincheng Lu, He Zhu, David Buckeridge, and Yue Li. TrajGPT: Irregular time-series representation learning of health trajectory. *IEEE J. Biomed. Health Inform.*, PP:1–14, October 2025.

Clark Mingxuan Ju, Liam Collins, Leonardo Neves, Bhuvesh Kumar, Louis Yufeng Wang, Tong Zhao, and Neil Shah. Generative recommendation with semantic ids: A practitioner's handbook. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, 2025.

Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, Ed Chi, and Xinyang Yi. Better generalization with semantic ids: A case study in ranking for recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 1039–1044, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688190. URL https://doi.org/10.1145/3640457.3688190.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Keshavan, Trung Vu, Lukasz Heidt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11513–11522, 2022. doi: 10.1109/CVPR52688.2022.01123.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1):160035, May 2016.

Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Celi, Roger Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 09 2018. doi: 10.1038/sdata.2018.178.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6341–6350. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7213-poincare-embeddings-for-learning-hierarchical-representations.pdf.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82, Nov 2024. ISSN 2662-8449. doi: 10.1038/s43586-024-00363-x. URL https://doi.org/10.1038/s43586-024-00363-x.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline (see above), or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are that SMI incorporates hierarchical semantics for better prediction, enhanced explainability, and cross-site robustness, which are supported by experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the Conclusion section, including the lack of more fine-grained clinical concepts and the need of a broader range of clinical prediction tasks.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not introduce new theoretical framework.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper ensures reproducibility by detailing the public datasets (MIMIC-III, eICU) and preprocessing in Section 4. It also provides the specific experimental settings, evaluation metrics, and hyperparameters for each task.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper utilizes publicly available and cited datasets (MIMIC-III and eICU).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 4, we provide details of the settings for the clinical prediction tasks, including the data split (80/20), optimizer (Adam), learning rate (1×10^{-4}), weight decay (1×10^{-5}), batch size (64), and the use of early stopping.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not include statistical significance analysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: This paper does not specify the computational resources

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that the research conforms.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper thoroughly discusses the positive societal impacts of the work, such as improving clinical prediction, clinical interpretability, and cross-site robustness.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks of releasing new datasets or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For the clinical datasets, we follow the required data use agreement (DUA) approvals from PhysioNet.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or direct interaction with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research did not involve new human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were used only for standard proofreading and editing, which does not require declaration per the guidelines.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.