SMI: Ontology-Grounded Semantic IDs for Hierarchy-Aware Medical Concept Representation

Ziyang Song[1,*], Lia Shen[2], Yixuan Li[3], Qincheng Lu[3]
Weihao Li[4], Haoran Xu[5], He Zhu[3], Xinyu Wang[3], Yue Li[3,*]

[1]School of Electrical Engineering and Computer Science, Ohio University
[2]Department of Quantitative and Computational Biology, University of Southern California
[3]School of Computer Science, McGill University
[4]School of Engineering, Northwestern University
[5]Department of Electrical & Computer Engineering, University of Washington
[*]Corresponding authors: ziyangs@ohio.edu, yueli@cs.mcgill.ca

Abstract

Biomedical language models (LMs) provide strong representations for clinical tasks, but they often fail to capture the hierarchical structure among medical concepts. This limitation is especially pronounced for rare and fine-grained clinical codes in long-tailed vocabularies, where hierarchy-agnostic LM representations become less informative as concept frequency decreases. We introduce Semantic Medical Identifier (SMI), a training-free framework for constructing ontology-grounded Semantic IDs and hierarchy-aware medical concept representations. For each diagnosis, procedure, or medication concept, SMI defines its Semantic ID as the coarse-to-fine path through an expert-defined medical ontology. SMI then aggregates frozen biomedical LM embeddings across the corresponding ontology levels to form hierarchy-aware medical concept representations that preserve coarse-to-fine clinical semantics. The base SMI uses uniform aggregation across ontology levels, while a level-weighted variant selects global level weights on validation data. Our theoretical analysis shows that SMI yields more informative medical concept representations than hierarchy-agnostic LM embeddings, particularly for rare and fine-grained clinical concepts. Experiments show that SMI captures meaningful hierarchical structure and improves clinical relation prediction. On MIMIC-IV, SMI improves mortality prediction, readmission prediction, drug recommendation, and rare drug prediction. On eICU, SMI yields more robust patient representations under cross-hospital distribution shift. The code is available on GitHub.

1 Introduction

The widespread adoption of electronic health record (EHR) systems has produced large-scale clinical data, enabling broad applications of language models (LMs) in healthcare. Biomedical LMs learn generalizable clinical representations for downstream tasks such as diagnosis prediction, drug recommendation, and patient risk stratification [Song et al., 2024, 2025, Wang et al., 2026]. However, most biomedical LMs remain predominantly data-driven and struggle to capture the hierarchical structure among medical concepts. Clinical coding systems such as the International Classification of Diseases (ICD) and the Anatomical Therapeutic Chemical (ATC) organize diagnoses, procedures, and medications into multi-level taxonomies, ranging from broad categories to fine-grained codes [Steindel, 2010, Chen et al., 2012]. Without explicitly modeling these medical ontologies, hierarchy-agnostic LMs encode each medical concept largely from its single-level textual description [Gao et al., 2019], limiting their ability to preserve coarse-to-fine clinical semantics.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

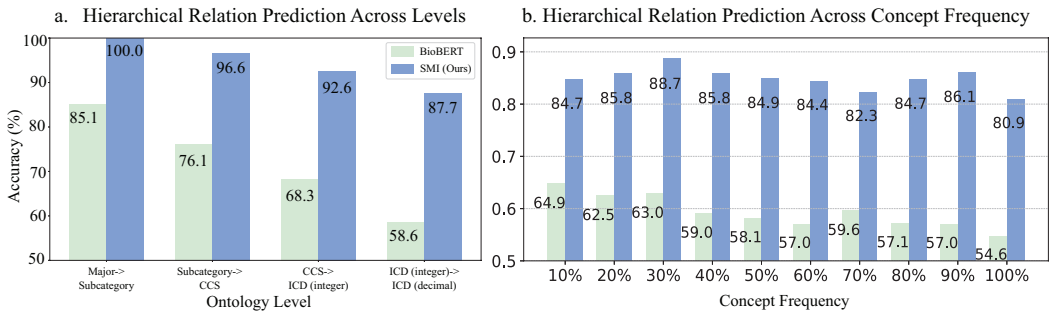

Figure 1: **SMI better distinguishes hierarchical relations than BioBERT on ICD 1-hop prediction.** (a) SMI consistently improves hierarchical relation prediction across levels, with the largest gains at finer ICD granularities. (b) SMI maintains stronger and more stable performance than BioBERT across concept-frequency bins, suggesting that medical ontology structure helps mitigate representation degradation for long-tailed clinical codes.

To demonstrate this limitation, we evaluated ICD 1-hop prediction using BioBERT, a common medical LM pre-trained on PubMed abstracts and PMC articles [Lee et al., 2020], and our Semantic Medical Identifier (SMI). This task evaluates whether a child concept is embedded closer to its true parent than to unrelated concepts. As shown in Fig. 1a, BioBERT performs reasonably well at the coarsest levels (85.1%), but its accuracy deteriorates rapidly as the hierarchy becomes finer, dropping to 58.6% at the finest level. The drop is most severe for fine-grained concepts, which are far more numerous and rarer in EHR data, making their embeddings harder to learn reliably (Fig. 7). To further examine this long-tail effect, we stratified fine-grained ICD concepts by frequency and evaluated 1-hop prediction. As shown in Fig. 1b, BioBERT accuracy decreases as concept frequency decreases, indicating that hierarchy-agnostic LM embeddings are vulnerable to long-tailed clinical code distributions. In contrast, SMI maintains strong and stable accuracy across frequency bins, suggesting that medical ontology structure can help mitigate representation degradation for infrequent fine-grained concepts. These observations motivate leveraging expert-defined medical ontologies to improve representations of long-tailed clinical codes under data scarcity.

Prior structure-aware approaches for medical concept representation broadly fall into two categories. The first learns structure-aware concept embeddings by incorporating curated biomedical resources such as UMLS [Liu et al., 2021], knowledge graphs [Sakhovskiy et al., 2024], and ontologies [He et al., 2024], but usually requires additional structure-aware training. The second develops task-specific prediction models that use ontology structure [Choi et al., 2017, Shang et al., 2019, Tan et al., 2025], but these methods are optimized for specific tasks rather than general-purpose concept representation across clinical applications. These limitations leave a gap for a hierarchy-aware framework that directly constructs concept representations without training an explicit ontology encoder or making representation construction dependent on a particular downstream task. SMI addresses this gap by directly aggregating frozen, level-specific LM embeddings along expert-defined medical ontology paths to construct hierarchy-aware medical concept representations without additional training.

In this study, we introduce SMI, a training-free framework for constructing ontology-grounded Semantic IDs and hierarchy-aware medical concept representations. SMI derives Semantic IDs from expert-defined medical ontologies. Diagnosis and procedure concepts are represented using coarse-to-fine ICD taxonomy, while medication concepts follow the ATC taxonomy. SMI then aggregates frozen biomedical LM embeddings along each Semantic ID to form a hierarchy-aware medical concept representation that preserves coarse-to-fine clinical semantics. The base SMI assigns equal importance to all hierarchy levels, while a level-weighted variant accounts for their different contributions by selecting a global weight for each level on validation data. Theoretically, we show that SMI yields more informative medical concept representations than hierarchy-agnostic LM embeddings, particularly for rare and fine-grained clinical concepts. Empirically, clinical relation prediction experiments show that SMI produces concept representations that better align with medical ontology structure, improving relation prediction within ICD and ATC as well as cross-ontology ICD-to-ATC prediction. On the MIMIC-IV dataset, SMI improves mortality prediction, readmission prediction, and drug recommendation, with particularly strong gains in rare drug prediction. On the eICU dataset, SMI shows significantly smaller cross-hospital distribution shifts than BioBERT, indicating improved cross-site robustness.

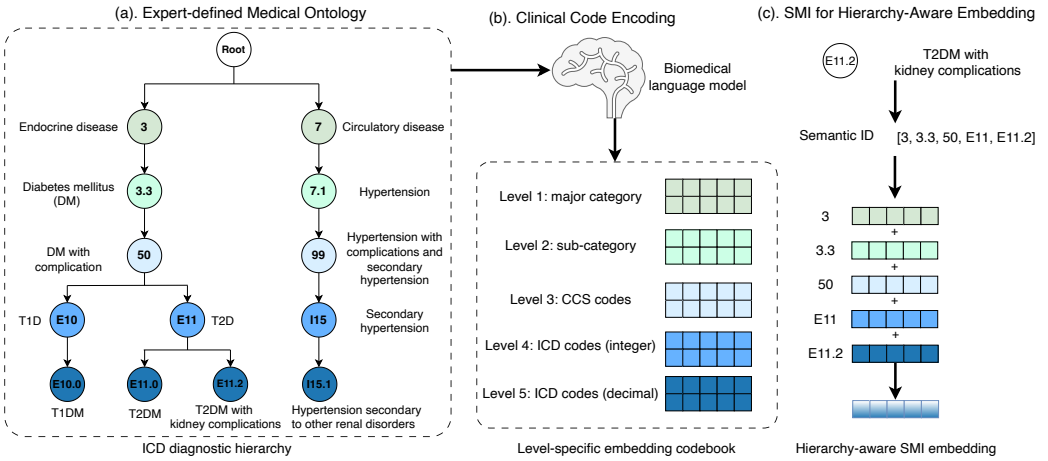

Figure 2: Overview of SMI and hierarchy-aware medical concept embedding. **a**. An expert-curated medical ontology organizes clinical concepts from coarse to fine levels. **b**. A frozen biomedical LM encodes clinical codes at each ontology level into level-specific embeddings. **c**. SMI represents each medical concept by its coarse-to-fine ontology path as a Semantic ID and aggregates the corresponding embeddings to form a hierarchy-aware medical concept representation.

2 Related Work

Prior work on biomedical concept representation with external knowledge mainly follows two directions. One direction learns structure-aware concept embeddings by injecting curated biomedical resources into representation learning. Existing methods use UMLS synonym sets [Sung et al., 2020, Liu et al., 2021], ontology definitions and descriptions [Remy et al., 2022, 2024], biomedical knowledge graphs [Zou et al., 2022, Yuan et al., 2022, Sakhovskiy et al., 2024], and medical ontologies [He et al., 2024, Yang et al., 2025]. While these methods improve concept embeddings using external biomedical knowledge, they typically require additional training rather than the direct use of frozen LM representations. The other direction develops supervised downstream prediction models that incorporate ontology structure through ontology-guided aggregation [Choi et al., 2017], graph-augmented patient modeling [Shang et al., 2019], expert-guided topic modeling [Song et al., 2022, Yang et al., 2026], or structure-semantics fusion [Tan et al., 2025]. These methods are trained for specific predictive tasks, whereas SMI constructs task-agnostic concept representations for downstream clinical applications. In contrast, SMI uses expert-defined ontology paths to construct hierarchy-aware concept representations from frozen LM embeddings without training.

Semantic IDs represent entities as structured sequences of discrete codes, allowing semantically related entities to share common prefixes and capture coarse-to-fine semantic structure. TIGER constructs hierarchical Semantic IDs by applying a residual-quantized variational autoencoder (RQ-VAE) to item content embeddings Rajput et al. [2023]. Subsequent work further improves learned Semantic IDs by aligning language and collaborative information Zheng et al. [2024], incorporating collaborative signals into item tokenization [Wang et al., 2024], and jointly learning item tokens with the downstream model Liu et al. [2025]. These methods construct discrete identifiers through learned quantization or tokenization. In contrast, SMI derives ontology-grounded Semantic IDs directly from expert-defined medical ontologies, with each Semantic ID corresponding to a medical concept's coarse-to-fine ontology path. SMI then aggregates frozen biomedical LM embeddings along these ontology paths to construct hierarchy-aware medical concept representations without training an additional quantizer or tokenization model.

3 Methodology

This section presents the SMI framework for hierarchy-aware concept representation of clinical codes, including diagnoses, procedures, and medications. Section 3.1 describes the construction of medical ontologies. Section 3.2 introduces the SMI framework. Section 3.3 provides a theoretical analysis showing that SMI yields more informative representations than hierarchy-agnostic LMs.

3.1 Expert-Curated Medical Ontologies

Clinical coding systems, such as ICD and ATC, organize diagnoses, procedures, and medications through multi-level taxonomies ranging from broad categories to fine-grained conditions. We organize clinical concepts into distinct five-level hierarchies, where each code has a unique path. For diagnostic codes, we adopt the Clinical Classifications Software (CCS) diagnostic taxonomy and construct a hierarchy of Major category $\rightarrow$ Sub-category $\rightarrow$ CCS code $\rightarrow$ ICD-CM code (integer) $\rightarrow$ ICD-CM code (decimal). We use the mutually exclusive AHRQ CCS mapping, assigning each full ICD-CM code to a single CCS category and a unique ontology path. For procedure codes, we follow the CCS procedure taxonomy and construct a five-level hierarchy: major category, sub-category, CCS code, prefix ICD-PCS code (first three characters), and full ICD-PCS code. For medications, we leverage the inherent five-level hierarchy of the ATC system: Anatomical Main Group $\rightarrow$ Therapeutic Subgroup $\rightarrow$ Pharmacological Subgroup $\rightarrow$ Chemical Subgroup $\rightarrow$ Chemical Substance. We provide the construction procedure and key statistics of medical ontologies in Appendix C.

3.2 SMI: Semantic IDs for Medical Concept Representation

Clinical concepts are organized in an expert-curated medical ontology with L granularity levels. Let $C^{(\ell)} = \{c_1^{(\ell)}, \ldots, c_{|C^{(\ell)}|}^{(\ell)}\}$ denote the set of clinical codes at level $\ell \in \{1, \ldots, L\}$. As illustrated in Fig. 2, the *SMI* of concept i is its ontology path in the hierarchy, defined as the ordered sequence of ancestor codes from coarse to fine levels:

$$\mathcal{P}(i) \equiv \big(a_i^{(1)}, \ldots, a_i^{(L)}\big) \tag{1}$$

where $a_i^{(\ell)} \in C^{(\ell)}$ denotes the ancestor of concept i at level ℓ. As shown in Fig. 2c, the ICD-10-CM code `E11.2` (type 2 diabetes mellitus with kidney complications) follows the path `[3, 3.3, 50, E11, E11.2]` corresponding to endocrine disease (Major category), diabetes mellitus (Sub-category), diabetes mellitus with complication (CCS code), type 2 diabetes mellitus (integer-level ICD), and itself (decimal-level ICD).

We construct a *hierarchy-aware medical concept representation* for concept i by aggregating level-specific embeddings of the codes in its Semantic ID $\mathcal{P}(i)$. For each clinical code $c_j^{(\ell)}$ at level ℓ, we BioBERT as a frozen biomedical LM encoder $\mathrm{Enc}(\cdot)$ to encode its textual description and then obtain its embedding by mean pooling the token embeddings:

$$\mathbf{e}_j^{(\ell)} = \mathrm{meanpool}\big(\mathrm{Enc}(c_j^{(\ell)})\big) \tag{2}$$

Let $\mathbf{E}^{(\ell)} \in \mathbb{R}^{|C^{(\ell)}| \times d}$ denote the level-ℓ embedding table, whose row $\mathbf{e}_j^{(\ell)} \in \mathbb{R}^d$ represents code $c_j^{(\ell)}$. We obtain the SMI embedding of concept i by summing the level-specific embeddings along $\mathcal{P}(i)$:

$$\mathbf{e}_i^{\text{SMI}} = \sum_{\ell=1}^{L} \mathbf{e}_{a_i^{(\ell)}}^{(\ell)} \tag{3}$$

This yields a d-dimensional SMI embedding $\mathbf{e}_i^{\text{SMI}}$ that preserves coarse-to-fine semantics without introducing additional trainable parameters.

The base SMI in equation 3 uses uniform aggregation across hierarchy levels. To account for their different contributions to the final representation, we consider a level-weighted variant, SMI-W, with *global* level weights shared across all concepts. Let $w_1, \ldots, w_L$ denote the level weights, where $w_\ell \geq 0, \sum_{\ell=1}^{L} w_\ell = L$. We define the weighted SMI representation of concept i as

$$\mathbf{e}_i^{\text{SMI-W}} = \sum_{\ell=1}^{L} w_\ell \, \mathbf{e}_{a_i^{(\ell)}}^{(\ell)}. \tag{4}$$

SMI-W replaces uniform aggregation with global level weighting. When $w_\ell = 1$ for all ℓ, it reduces to the original SMI. The weights are shared across all concepts and selected on the validation set.

3.3 Theoretical Analysis of LM Representation Collapse and SMI Informativeness

We consider two embedding paradigms for clinical concepts. A *hierarchy-agnostic LM embedding* $\mathbf{e}_i^{\text{LM}}$ is obtained by directly encoding the textual description of concept i with a biomedical LM. In

contrast, our proposed *hierarchy-aware SMI embedding* $\mathbf{e}_i^{\mathrm{SMI}}$ aggregates embeddings of all ancestor concepts along its hierarchical path. Both aim to estimate the latent semantic vector $\boldsymbol{\mu}_i$. However, $\mathbf{e}_i^{\mathrm{LM}}$ depends only on data-driven textual signal, whereas $\mathbf{e}_i^{\mathrm{SMI}}$ leverages higher-frequency ancestor information to reduce estimation noise variance and thus improves representations. All proofs in this section are detailed in Appendix B.

Prior work has shown that word frequency affects contextualized embedding quality in LMs [Schick and Schütze, 2020a,b], which is consistent with our observation in Fig. 1b that BioBERT accuracy declines for less frequent concepts. Motivated by these findings, we model hierarchy-agnostic LM embeddings $\mathbf{e}_i^{\mathrm{LM}}$ as frequency-dependent noisy estimates of latent concept semantics $\boldsymbol{\mu}_i$, with estimation variance decreasing as frequency n_i increases. Because clinical vocabularies are long-tailed (Table 7), many fine-grained concepts have small n_i and thus noise-dominated embeddings. This motivates the following frequency-dependent estimation assumption.

Assumption 3.1 (Frequency-dependent embedding estimation)**.** For each concept i, let $\boldsymbol{\mu}_i \in \mathbb{R}^d$ denote its true semantic vector and n_i its frequency in the training corpus. We estimate its LM embedding $\mathbf{e}_i^{\mathrm{LM}}$ as a true signal perturbed by a zero-mean isotropic noise:

$$\mathbf{e}_i^{\mathrm{LM}} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{n_i}\mathbf{I}_d\right).$$

Assumption 3.1 implies that LM embedding noise decreases with concept frequency. In long-tailed clinical vocabularies, fine-grained concepts typically occur infrequently and thus suffer from high noise. This motivates a signal-to-noise (SNR) criterion for measuring semantic separability.

Definition 3.2 (Pairwise SNR)**.** Given two concepts i and j with true semantic vectors $(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$ and noises $(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j)$, we define the pairwise SNR criterion as the ratio of their semantic distance to the expected noise magnitude:

$$\mathrm{SNR}(i,j) := \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sqrt{\mathbb{E}\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|^2}}.$$

Under Lemma B.2, this expression simplifies to

$$\mathrm{SNR}(i,j) = \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma\sqrt{d\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}.$$

Under Assumption 3.1 and the long-tailed frequency distribution, fine-grained concepts exhibit vanishing pairwise SNR, leading to the following theorem:

Theorem 3.3 (Inseparability of infrequent concepts with hierarchy-agnostic LM embeddings)**.** *Let i and j be two fine-grained concepts with low frequencies. The semantic difference between the two concepts becomes negligible relative to the noise:*

$$\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|} \xrightarrow{p} 0,$$

where $\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|$ represents the magnitude of noise difference between the two concepts.

Theorem 3.3 shows that, in long-tailed clinical vocabularies, hierarchy-agnostic LMs produce noise-dominated embeddings for infrequent fine-grained concepts, causing semantic separability to degrade at finer hierarchy levels.

Remark 3.4. Pairwise distances between infrequent fine-grained concepts concentrate in a narrow "noise shell," making their representations geometrically indistinguishable. Thus, 1-hop prediction with hierarchy-agnostic LM embeddings degrades to near-random results (Fig. 1a).

Inspired by the additive compositionality of word embeddings [Mikolov et al., 2013], we model a fine-grained concept as the sum of its coarse-grained parent semantics and a child-specific semantic offset. By leveraging more informative parent representations, SMI mitigates the data scarcity and noise of fine-grained concepts. We formalize this hierarchical semantic decomposition as follows:

Assumption 3.5 (Hierarchical semantic decomposition)**.** For a child concept c and its parent concept p, we assume the child's semantic vector decomposes as

$$\boldsymbol{\mu}_c^{(\ell+1)} = \boldsymbol{\mu}_p^{(\ell)} + \boldsymbol{\delta}_c^{(\ell+1)},$$

where $\boldsymbol{\mu}_p^{(\ell)}$ captures the shared coarse-grained semantics of the parent concept and $\delta_c^{(\ell+1)}$ captures the child-specific semantic offset (e.g., severity or complication of a disease). We assume offsets are zero-centered ($\mathbb{E}[\delta] = 0$) and are independent across child concepts.

Appendix B.6 provides empirical offset consistency analysis supporting this decomposition using real ICD-9 parent-child concept pairs. Offsets sharing the same clinical modifier are substantially more aligned than offsets from different modifier groups, with mean cosine similarity 0.655 versus 0.080 ($p = 2.0 \times 10^{-4}$). Under Assumption 3.5, SMI improves the representation of infrequent fine-grained concepts by anchoring them to higher-frequency ancestor semantics with lower variance, rather than relying only on infrequent fine-grained codes. This explains why SMI can mitigate representation collapse in long-tailed clinical vocabularies. The following proposition formalizes this effect by showing that SMI preserves separability between an infrequent concept and non-parent concepts.

Proposition 3.6 (SMI preserves separability against non-parent concepts)**.** *Under Assumptions 3.1 and 3.5, let c be a child concept with parent p, and let $j \neq p$ be any other concept. If $\|\mu_p - \mu_j\| > 0$ then as $n_c \to 0$,*

$$\frac{\|\mu_c^{\mathrm{SMI}} - \mu_j^{\mathrm{SMI}}\|}{\|\epsilon_c^{\mathrm{SMI}} - \epsilon_j^{\mathrm{SMI}}\|} \xrightarrow{p} \frac{\|\mu_p - \mu_j\|}{\|\epsilon_p - \epsilon_j\|} > 0.$$

Proposition 3.6 shows that SMI embeddings are separable from non-parent concepts even for infrequent fine-grained concepts, whereas Theorem 3.3 indicates that hierarchy-agnostic LM embeddings are inseparable. Although both approaches estimate the same latent semantics, SMI improves representations by reducing embedding variance rather than increasing semantic margins. By aggregating higher-frequency ancestor embeddings, SMI reduces noise in infrequent fine-grained concepts. This leads to the following theorem, which establishes that SMI achieves strictly higher SNR than hierarchy-agnostic LM embeddings by reducing estimation noise variance.

Theorem 3.7 (Weighted SMI Improves SNR via Variance Reduction)**.** *Under Assumption 3.1, suppose the level-wise estimation noises are independent and at least one ancestor level is more frequent than the leaf concept, i.e., $n_i^{(\ell)} > n_i^{(L)}$ for some $\ell < L$. After normalizing the hierarchy-level weights, weighted SMI strictly reduces estimation noise variance of infrequent fine-grained concepts:*

$$\mathbb{E}\,\|\boldsymbol{\epsilon}_i^{\text{SMI-W}}\|^2 < \mathbb{E}\,\|\boldsymbol{\epsilon}_i^{\mathrm{LM}}\|^2.$$

Consequently, for any pair (i, j) with fixed semantic margin $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$, the pairwise SNR satisfies

$$\mathrm{SNR}_{\text{SMI-W}}(i, j) > \mathrm{SNR}_{\mathrm{LM}}(i, j).$$

Corollary 3.8 (Uniform SMI improves SNR)**.** *Uniform SMI is a special case of weighted SMI with equal level weights. Therefore, under the same conditions as Theorem 3.7, uniform SMI strictly reduces estimation noise variance after normalizing the equal hierarchy-level weights:*

$$\mathbb{E}\,\|\boldsymbol{\epsilon}_i^{\mathrm{SMI}}\|^2 < \mathbb{E}\,\|\boldsymbol{\epsilon}_i^{\mathrm{LM}}\|^2.$$

Uniform SMI also achieves higher pairwise SNR than the hierarchy-agnostic LM embedding:

$$\mathrm{SNR}_{\mathrm{SMI}}(i, j) > \mathrm{SNR}_{\mathrm{LM}}(i, j).$$

Uniform SMI is a special case of weighted SMI with equal hierarchy-level weights. Therefore, under the same conditions as Theorem 3.7, uniform SMI also strictly reduces estimation-noise

4 Experiments and Results

4.1 SMI Embeddings Capture Clinical Concept Relations

To assess whether SMI embeddings capture clinical relations in medical ontologies, we evaluated multi-hop relation prediction within the ICD and ATC hierarchies by predicting ancestor–descendant links within each hierarchy. We also evaluated a cross-ontology 1-hop relation prediction between treatments and their corresponding diagnoses (ICD→ATC). The experimental setup is provided in Appendix E.2.

We compared SMI with three groups of baselines: (1) the hierarchy-agnostic LM backbone BioBERT [Lee et al., 2020]; (2) LMs with structure knowledge, including SapBERT [Liu et al.,

Table 1: Clinical relation prediction results for ICD→ICD, ATC→ATC, and ICD→ATC, comparing SMI and SMI-W against baselines. Best F-scores are highlighted in bold. We reported the optimized result for each method as the decision threshold is selected via deterministic grid search.

Method	ICD			ATC			ICD→ATC		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
SMI	70.6	67.9	69.2	83.8	67.6	74.9	53.2	61.4	57.0
SMI-W	77.2	66.6	**71.5**	86.1	70.5	**77.5**	53.6	62.2	**57.6**
BioBERT	30.4	75.8	44.2	50.7	45.7	48.1	49.3	57.0	50.9
SapBERT	69.9	64.5	67.1	84.0	66.9	74.5	50.2	55.7	52.9
HiTs	69.3	61.7	65.3	83.0	66.2	71.5	50.5	56.3	53.2
OnT	65.9	59.1	62.3	78.5	62.3	69.5	48.9	55.2	51.9
BERGAMOT	66.5	70.2	55.3	79.4	69.7	74.2	47.9	56.2	51.7
RotatE	69.3	58.2	63.3	80.3	63.9	71.2	49.2	57.2	52.9
cui2vec	68.3	60.3	64.1	77.2	61.4	68.4	45.7	54.0	49.5

Table 2: Aggregate clustering quality of ICD diagnosis embeddings by major disease category. NMI, ARI, Purity, Recall, Precision, and Silhouette are higher better, while Intra/Inter Ratio is lower better. SMI achieves stronger cluster-label alignment and better geometric separation than BioBERT, indicating more ontology-consistent diagnosis embeddings.

Method	NMI ↑	ARI ↑	Purity ↑	Recall ↑	Precision ↑	Sil. ↑	Intra/Inter ↓
SMI	**0.77**	**0.56**	**0.78**	**0.80**	**0.61**	**0.52**	**0.29**
BioBERT	0.39	0.25	0.51	0.41	0.35	0.04	0.82

2021], HiTs [He et al., 2024], OnT [Yang et al., 2025], and BERGAMOT [Sakhovskiy et al., 2024]; and (3) structure-aware embedding baselines, including RotatE [Sun et al., 2019] and cui2vec [Beam et al., 2020]. Implementation details are provided in Appendix D.4.

As shown in Table 1, SMI and SMI-W achieve the best performance across all three relation types, outperforming all baseline groups. For within-ontology prediction, SMI-W obtains the highest F-scores on ICD (71.5%) and ATC (77.5%), substantially improving over the strongest baseline SapBERT on ICD (67.1%) and ATC (74.5%). This indicates that weighting hierarchy levels helps capture parent-child relationships more effectively. Cross-ontology prediction (ICD→ATC) is challenging because methods trained within each ontology may not directly align diagnosis and medication semantics. Despite this, SMI-W achieves the best Precision (53.6%), Recall (62.2%), and F-score (57.6%), outperforming the strongest baseline HiTs with F-score 53.2%. These results suggest that aggregating frozen biomedical LM embeddings along ontology-grounded Semantic IDs improves medical concept representations for both within-ontology and cross-ontology relation prediction tasks.

4.2 SMI Embeddings Capture Medical Ontology Structure

We applied the Uniform Manifold Approximation and Projection (UMAP) to visualize ICD diagnosis embeddings in 2D [McInnes et al., 2018]. Each point represents a decimal-level ICD code with frequency ≥ 5 and is colored by its major disease category. As shown in Fig. 3, BioBERT embeddings are highly mixed across major categories, indicating that hierarchy-agnostic LM embeddings do not clearly separate disease groups. In contrast, SMI embeddings form more distinct clusters by major disease category and further split into coherent subgroups that align with subcategories in the medical ontology. This comparison suggests that hierarchy-aware aggregation helps SMI capture clinically meaningful structure that is not evident from hierarchy-agnostic LM embeddings.

To quantify whether the embeddings align with the ICD ontology, we applied K-means clustering with the number of clusters set to the number of major disease categories. As shown in Table 10, SMI achieves stronger ontology-consistent clustering than BioBERT, with higher NMI (0.77 vs. 0.39), ARI (0.56 vs. 0.25), and Purity (0.78 vs. 0.51). SMI also improves per-category clustering quality, increasing Recall from 0.41 to 0.80 and Precision from 0.35 to 0.61, indicating more coherent grouping within the same major category. In addition, SMI better separates ICD major categories,

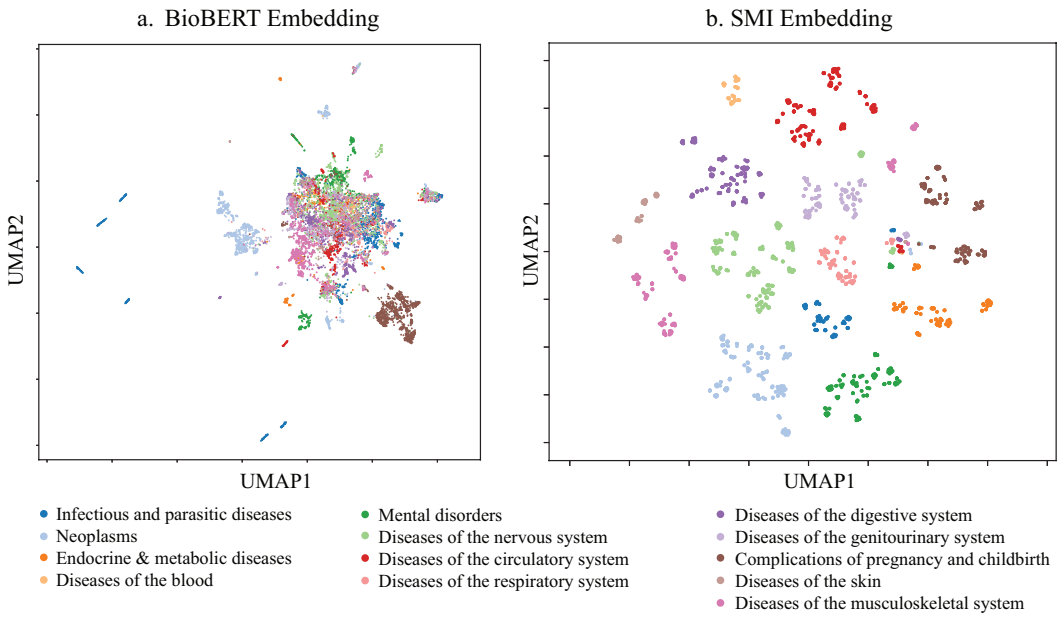

Figure 3: Visualization of ICD diagnosis embeddings across 13 major categories using UMAP. (a) BioBERT embeddings. (b) SMI embeddings. ICD codes embedded by SMI form more distinct category-level clusters that are better aligned with the medical ontology.

with a higher Silhouette score (0.52 vs. 0.04) and a lower Intra/Inter Ratio (0.29 vs. 0.82). Full aggregate and per-category clustering results with detailed analyses are provided in Appendix E.3.

4.3 Evaluation on MIMIC-IV and eICU EHR Datasets

We evaluated SMI medical concept representations on two EHR datasets, MIMIC-IV and eICU Johnson et al. [2023], Pollard et al. [2018]. As the first EHR dataset, we used MIMIC-IV for clinical prediction tasks, including mortality and readmission prediction as well as drug recommendation. We extracted clinical codes from 123,488 patients and mapped them to their corresponding five-level Semantic IDs. As the second EHR dataset, we used the eICU dataset to evaluate distribution shifts in patient observations across hospitals using both SMI and BioBERT embedding methods. We mapped diagnoses and medications to 72 unique integer-level ICD-9 codes and 241 unique five-level ATC codes. We selected the 10 hospitals with the largest number of patients, resulting in a total of 24,179 patients and an average of 2,418 patients per hospital. We split each dataset into training, validation, and test sets using a 70%/10%/20% split. Detailed settings for the MIMIC-IV clinical prediction tasks and the eICU distribution shift analysis are provided in Sections 4.4 and 4.5, respectively.

4.4 SMI Improves Clinical Prediction and Rare Drug Prediction

For clinical prediction, we extracted diagnosis, procedure, and medication codes from MIMIC-IV and evaluated three common tasks. Mortality prediction assesses whether a patient has died within 90 days after discharge. Readmission prediction assesses whether a patient is readmitted within the next 15 days after discharge. Drug recommendation predicts the set of medications for each visit given the patient's prior clinical history. Mortality and readmission are binary classification tasks evaluated with Area Under Precision-Recall Curve (AUPRC) and Area Under Receiver Operating Characteristic Curve (AUROC). Drug recommendation is a multi-label prediction task evaluated with AUPRC, F-score, and Jaccard. We also predicted rare drugs using 98 ATC codes with frequency between 10 and 20, using Recall of top 15 predictions as a metric. We mapped clinical codes to their SMI embeddings and mean-pooled the code embeddings to construct patient-level representations. Mortality and readmission use a linear binary classification head, while drug recommendation uses a linear multi-label classification head. We compared with two groups of baselines: (1) clinical concept embedding methods such as GRAM [Choi et al., 2017], cui2vec [Beam et al., 2020], and RotatE [Sun et al., 2019]; (2) biomedical LMs including BioBERT [Lee et al., 2020], SapBERT [Liu et al.,

Table 3: Clinical prediction results on MIMIC-IV for mortality, readmission, and (rare) drug recommendation, comparing SMI with two groups of baselines. Metrics are reported as mean (standard error) from bootstrap. The best results in each column are highlighted in bold.

Method	Mortality		Readmission		Drug			Rare Drug
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	F-score	Jaccard	Recall@15
SMI	44.5 (0.5)	87.4 (1.0)	**48.9 (0.2)**	72.5 (0.5)	75.6 (0.1)	62.4 (0.1)	48.6 (0.1)	41.4 (0.2)
SMI-W	**45.1 (0.3)**	**88.0 (0.7)**	48.5 (0.3)	**72.9 (0.3)**	**77.6 (0.1)**	**65.3 (0.2)**	**50.5 (0.1)**	**44.2 (0.3)**
GRAM	38.5 (0.7)	79.1 (0.9)	43.2 (0.8)	68.2 (0.4)	72.0 (0.3)	61.3 (0.4)	48.6 (0.4)	34.5 (0.2)
cui2vec	39.0 (0.4)	80.4 (0.8)	43.7 (0.6)	68.9 (0.3)	75.2 (0.1)	64.3 (0.3)	49.9 (0.4)	35.6 (0.1)
RotatE	39.2 (0.3)	79.6 (0.6)	43.6 (0.4)	69.2 (0.6)	74.1 (0.2)	62.7 (0.4)	48.7 (0.4)	36.4 (0.1)
HiTs	42.7 (0.6)	81.5 (0.7)	45.2 (0.4)	69.6 (0.8)	74.8 (0.2)	61.6 (0.2)	48.0 (0.1)	37.5 (0.1)
OnT	43.3 (0.5)	83.4 (1.1)	45.9 (0.7)	70.3 (0.8)	71.0 (0.1)	57.3 (0.2)	44.5 (0.1)	35.8 (0.1)
BioBERT	43.2 (0.9)	86.1 (1.3)	46.5 (0.9)	71.5 (1.2)	73.9 (0.1)	60.3 (0.2)	44.5 (0.2)	37.6 (0.2)
SapBERT	42.4 (0.6)	83.7 (1.2)	45.3 (0.7)	70.2 (1.1)	74.6 (0.1)	60.5 (0.2)	45.7(0.4)	40.0 (0.3)
G-BERT	42.7 (0.4)	85.0 (1.3)	47.3 (0.4)	70.1 (0.7)	75.4 (0.1)	61.4 (0.3)	45.5 (0.3)	40.7 (0.3)
BEHRT	42.0 (0.3)	84.3 (0.7)	44.5 (0.4)	69.0 (0.8)	74.3 (0.2)	60.2 (0.2)	45.7 (0.3)	37.5 (0.3)
ClinicalBERT	43.5 (0.9)	86.2 (0.8)	45.7 (0.6)	70.4 (0.6)	73.8 (0.2)	59.4(0.4)	44.0 (0.4)	38.4 (0.4)
BoxLM	41.7 (1.4)	83.5 (1.2)	43.8 (0.8)	69.1 (0.9)	73.6 (0.4)	59.1 (0.2)	43.3 (0.5)	35.9 (0.5)
BERGAMOT	43.8 (0.9)	85.9 (0.7)	45.6 (0.6)	71.0 (1.1)	75.2 (0.3)	61.7 (0.6)	48.1 (0.3)	38.6 (0.5)

Table 4: Ablations on clinical relation prediction comparing the LM baseline, SMI, and SMI-W.

Method	ICD			ATC			ICD→ATC		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
BioBERT	30.4	75.8	43.4	50.7	45.7	48.1	49.3	57.0	52.9
SMI	70.6	67.9	69.2	83.8	67.6	74.9	53.2	61.4	57.0
SMI-W	77.2	66.6	71.5	86.1	70.5	77.5	53.6	62.2	57.6

2021], G-BERT [Shang et al., 2019], BEHRT [Li et al., 2020], ClinicalBERT [Alsentzer et al., 2019], BoxLM [Tan et al., 2025], and BERGAMOT [Sakhovskiy et al., 2024]. We provided implementation details in Appendix D.4.

As shown in Table 3, SMI and SMI-W consistently improve clinical prediction on MIMIC-IV. SMI and SMI-W generally achieve stronger mortality and readmission prediction performance than the baselines in terms of AUPRC and AUROC. The gains are especially clear for drug recommendation, where SMI-W achieves the best AUPRC, F-score, and Jaccard scores (77.6%, 65.3%, and 50.5%), significantly outperforming the strongest baseline ($p = 0.015$). For rare drug prediction, SMI-W achieves the highest Recall@15 (44.2%), significantly outperforming the strongest baseline G-BERT ($p = 0.0081$), showing that SMI is particularly effective for long-tailed fine-grained codes. These results suggest that aggregating frozen biomedical LM embeddings along ontology-grounded Semantic IDs benefits downstream clinical prediction, particularly rare drug prediction. Compared with GRAM, which performs hierarchy-aware aggregation within a supervised prediction model, SMI constructs hierarchy-aware medical concept representations from frozen biomedical LM embeddings without task-specific training. Compared with structure-aware encoders such as HiTs and OnT, which enforce structural consistency for individual concept embeddings, SMI explicitly aggregates hierarchy-level semantics into a final concept representation.

4.5 SMI Mitigates Cross-Hospital Distribution Shift

The multi-center eICU dataset exhibits substantial heterogeneity across hospitals in patient populations, clinical practice patterns, and data collection processes [Zhu et al., 2025]. We thus use *hospital* as the patient-group axis and evaluate cross-site distribution shift, as such heterogeneity can degrade the generalizability of learned patient representations. Specifically, we fit a hospital-specific VAE density model on patient embeddings encoded by either SMI or BioBERT [Kingma and Welling, 2013]. For a source hospital k and target hospital t, we measured the shift by the log-likelihood gap:

$$\Delta = \mathbb{E}_{x_k}\left[\log p_k(x_k)\right] - \mathbb{E}_{x_t}\left[\log p_k(x_t)\right] \tag{5}$$

where $\log p_k(x_k)$ and $\log p_k(x_t)$ are the log-likelihoods of in-domain x_k and out-of-domain x_t evaluated by the VAE trained on hospital k. More evaluation details are provided in Appendix E.5. Fig. 9 visualizes the cross-hospital distribution shift $|\Delta|$ for each source–target hospital pair under a source-trained density model, where larger $|\Delta|$ indicates greater shift between the source and target distributions. SMI yields consistently smaller shift than BioBERT (50.77 vs. 74.06), indicating improved cross-site robustness. Fig. 4 shows cross-hospital distribution shift between BioBERT and SMI by the difference ($|\Delta$BioBERT $-$ $|\Delta|$SMI) across source–target hospital pairs. Most off-diagonal entries are red, indicating larger shifts under BioBERT and thus better cross-site robustness for SMI. We also computed p-values using the Wilcoxon signed-rank paired test over all off-diagonal hospital pairs. SMI achieves significantly smaller cross-hospital shifts ($p = 1.97 \times 10^{-5}$), showing that SMI improves cross-site robustness of representations.

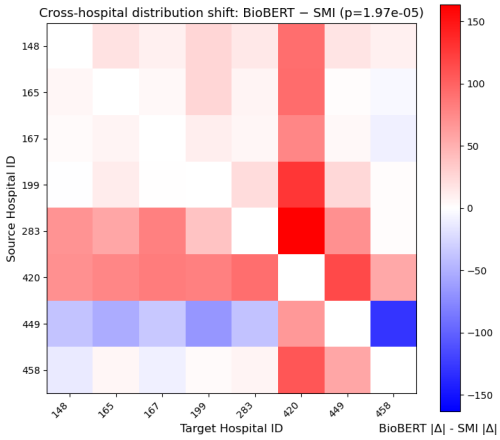

Figure 4: Cross-hospital distribution shift comparison between BioBERT and SMI. Each cell shows $|\Delta|_{\text{BioBERT}} - |\Delta|_{\text{SMI}}$; red colors indicate that SMI is more cross-site robust. P-value indicates that SMI achieves significantly lower distribution shift.

4.6 Ablation Studies

We ablate SMI on clinical relation prediction by comparing BioBERT, SMI, and SMI-W (Table 4). BioBERT directly encodes each code description without using ontology structure. As shown in Table 4, SMI consistently outperforms BioBERT across ICD, ATC, and ICD→ATC relation prediction, showing that simple aggregation of ontology path improves hierarchy-aware concept representation. SMI-W further improves over uniform SMI by selecting global level weights, suggesting that hierarchy levels contribute unequally to the final concept representation.

5 Conclusion

Our study makes the following key contributions. First, we propose SMI, a training-free framework that generates hierarchy-aware concept representations by integrating medical ontologies with LM embeddings. Second, we provide a theoretical analysis showing that SMI improves SNR and yields more informative embeddings. Third, qualitative analyses confirm that SMI learns clinically meaningful structures aligned with medical hierarchies. Finally, extensive experiments on clinical relation prediction, downstream clinical prediction, and cross-site distribution shift demonstrate the broad utility of SMI. The current study focuses on expert-curated coding systems such as ICD and ATC. Applying SMI to emerging disease classifications, newly introduced drugs, or local coding standards may require additional ontology validation and code mapping. For example, Canadian healthcare systems use ICD-10-CA or Drug Identification Numbers, which must be carefully mapped before deployment. Future work will extend SMI to clinician-validated ontology harmonization across coding systems, missing-link discovery, and broader biomedical knowledge resources. We will also study how to jointly integrate multiple ontologies within a unified SMI framework.

Acknowledgments

This work was supported in part by the NVIDIA Academic Grant Program through an award to Ziyang Song at Ohio University.

References

Ziyang Song, Qincheng Lu, Hao Xu, He Zhu, David Buckeridge, and Yue Li. Timelygpt: Extrapolatable transformer pre-training for long-term time-series forecasting in healthcare. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400713026. doi: 10.1145/3698587.3701364. URL https://doi.org/10.1145/3698587.3701364.

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.

Xinyu Wang, Hanwei Wu, Zhenghan Tai, Sicheng Lyu, Qincheng Lu, Ziyu Zhao, Jijun Chi, Jingrui Tian, Xiao-Wen Chang, and Ziyang Song. Saferx-agent: A knowledge-grounded multi-agent framework for safe and explainable medication recommendation, 2026. URL https://arxiv.org/abs/2605.29146.

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.

Lei Chen, Wei-Ming Zeng, Yu-Dong Cai, Kai-Yan Feng, and Kuo-Chen Chou. Predicting anatomical therapeutic chemical (atc) classification of drugs by integrating chemical-chemical interactions and similarities. *PLOS ONE*, 7(4):1–7, 04 2012. doi: 10.1371/journal.pone.0035254. URL https://doi.org/10.1371/journal.pone.0035254.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkEYojRqtm.

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.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*, pages 565–574, August 2021.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. Biomedical entity representation with graph-augmented multi-objective transformer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Yuan He, Moy Yuan, Jiaoyan Chen, and Ian Horrocks. Language models as hierarchy encoders. *Advances in Neural Information Processing Systems*, 37:14690–14711, 2024.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795, 2017.

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

Yanchao Tan, Hang Lv, Yunfei Zhan, Guofang Ma, Bo Xiong, and Carl Yang. BoxLM: Unifying structures and semantics of medical concepts for diagnosis prediction in healthcare. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=jIci8mGVeh.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. Biomedical entity representations with synonym marginalization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.335. URL https://aclanthology.org/2020.acl-main.335/.

François Remy, Kris Demuynck, and Thomas Demeester. BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1454–1465, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.104. URL https://aclanthology.org/2022.findings-emnlp.104/.

François Remy, Kris Demuynck, and Thomas Demeester. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *J. Am. Med. Inform. Assoc.*, 31(9):1844–1855, September 2024.

Yuesong Zou, Ahmad Pesaranghader, Ziyang Song, Aman Verma, David L Buckeridge, and Yue Li. Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. *Sci. Rep.*, 12(1):17868, October 2022.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *J. Biomed. Inform.*, 126 (103983):103983, February 2022.

Hui Yang, Jiaoyan Chen, Yuan He, Yongsheng Gao, and Ian Horrocks. Language models as ontology encoders. In *The Semantic Web – ISWC 2025: 24th International Semantic Web Conference, Nara, Japan, November 2–6, 2025, Proceedings, Part I*, page 443–461, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-09526-8. doi: 10.1007/978-3-032-09527-5_24. URL https://doi.org/10.1007/978-3-032-09527-5_24.

Ziyang Song, Yuanyi Hu, Aman Verma, David L. Buckeridge, and Yue Li. Automatic phenotyping by a seed-guided topic model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4713–4723, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3542675. URL https://doi.org/10.1145/3534678.3542675.

Ziqi Yang, Ziyang Song, Shadi Zabad, Marc-André Legault, and Yue Li. PheCode-guided multimodal topic modeling of electronic health records improves disease incidence prediction and GWAS discovery from UK biobank. *Brief. Bioinform.*, 27(1), January 2026.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BJ0fQUU32w.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, May 2024.

Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2400–2409, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679569. URL https://doi.org/10.1145/3627673.3679569.

Enze Liu, Bowen Zheng, Cheng Ling, Lantao Hu, Han Li, and Wayne Xin Zhao. Generative recommender with end-to-end learnable item tokenization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 729–739, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3729989. URL https://doi.org/10.1145/3726302.3729989.

Timo Schick and Hinrich Schütze. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8766–8774, Apr. 2020a. doi: 10.1609/aaai.v34i05.6403. URL https://ojs.aaai.org/index.php/AAAI/article/view/6403.

Timo Schick and Hinrich Schütze. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.368. URL https://aclanthology.org/2020.acl-main.368/.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

Andrew L Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 25, page 295, 2020.

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.

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

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.

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.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78, 2019.

He Zhu, Jun Bai, Na Li, Xiaoxiao Li, Dianbo Liu, David L Buckeridge, and Yue Li. Fedweight: mitigating covariate shift of federated learning on electronic health records data through patients re-weighting. *npj Digital Medicine*, 8(1):286, 2025.

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

Fabrício Kury. Scrape anatomical-therapeutic-chemical (atc) classes from the who collaborating centre for drug statistics methodology website, 2024. URL https://github.com/fabkury/atcd.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 2019.

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.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1646–1655. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ganea18a.html.

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

A Notation Table

Table 5: Summary of Notations and Variables used in SMI

Notation	Description
	SMI and Embedding
L	Total number of hierarchy levels
l	Index of a specific hierarchy level
$C^{(\ell)}$	Set of clinical codes at hierarchy level ℓ
$c_j^{(\ell)}$	The j-th clinical code at hierarchy level ℓ
$\mathcal{P}(i)$	Semantic ID of concept i
$a_i^{(l)}$	Ancestor code of concept i at level l
$\mathbf{e}_j^{(\ell)}$	Embedding of the j-th code at level ℓ
$\mathbf{E}^{(\ell)}$	Embedding matrix for level ℓ
$\text{Enc}(\cdot)$	Frozen biomedical LM encoder
$\mathbf{e}_i^{\text{SMI}}$	SMI embedding
w_ℓ	Global weight for hierarchy level ℓ
$\mathbf{e}_i^{\text{SMI-W}}$	Weighted SMI embedding
	Theoretical Analysis
$\mathbf{e}_i^{\text{LM}}$	Hierarchy-agnostic LM embedding
$\boldsymbol{\mu}_i$	True semantic vector for concept i
$\boldsymbol{\epsilon}_i$	Embedding estimation noise vector
n_i	Frequency of concept i in the training corpus
$\text{SNR}(i, j)$	Pairwise Signal-to-Noise Ratio between concepts i and j
$\delta_c^{(l+1)}$	Semantic offset of child concept c at level $\ell + 1$ with respect to its parent

B Theoretical Proofs and Empirical Support for Assumptions

B.1 Estimation Variance under Long-Tailed Concept Distribution (Assumption 3.1)

We model the learned embedding of a clinical concept i, denoted by $\mathbf{e}_i^{\mathrm{LM}} \in \mathbb{R}^d$, as a noisy estimate of its underlying true semantic vector $\boldsymbol{\mu}_i$:

$$\mathbf{e}_i^{\mathrm{LM}} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i.$$

Proposition B.1 (Estimator variance decreases with frequency $1/n_i$)**.** *Let $\mathbf{e}_i^{\mathrm{LM}}$ denote the empirical estimated embedding of concept i computed from n_i independent samples. We thus have:*

$$\mathbb{E}[\mathbf{e}_i^{\mathrm{LM}}] = \boldsymbol{\mu}_i, \qquad \mathbb{E}\left[\left\|\mathbf{e}_i^{\mathrm{LM}} - \boldsymbol{\mu}_i\right\|^2\right] = \frac{\sigma^2 d}{n_i}.$$

As the frequency n_i decreases, the embedding estimation becomes more unstable and its variance increases.

Proof. Let the empirical embedding of concept i be

$$\mathbf{e}_i^{\mathrm{LM}} = \frac{1}{n_i}\sum_{k=1}^{n_i} \mathbf{x}_k,$$

where each sample $\mathbf{x}_k \in \mathbb{R}^d$ satisfies $\mathbb{E}[\mathbf{x}_k] = \boldsymbol{\mu}_i$ and $\mathrm{Cov}(\mathbf{x}_k) = \Sigma_i$, independently across k.

We first show that the empirical mean is an unbiased estimator of the true semantic vector $\boldsymbol{\mu}_i$, i.e., $\mathbb{E}[\mathbf{e}_i^{\mathrm{LM}}] = \boldsymbol{\mu}_i$, which follows from the linearity of expectation:

$$\begin{aligned}
\mathbb{E}[\mathbf{e}_i^{\mathrm{LM}}] &= \mathbb{E}\left[\frac{1}{n_i}\sum_{i=1}^{n_i} \mathbf{x}_i\right] \\
&= \frac{1}{n_i}\sum_{k=1}^{n_i} \mathbb{E}[x_k] \\
&= \boldsymbol{\mu}_i.
\end{aligned}$$

Since the samples are independent, cross-covariance terms vanish, and the covariance is given by

$$\begin{aligned}
\mathrm{Cov}(\mathbf{e}_i^{\mathrm{LM}}) &= \mathrm{Cov}\left(\frac{1}{n_i}\sum_{k=1}^{n_i} \mathbf{x}_k\right) \\
&= \frac{1}{n_i^2}\sum_{k=1}^{n_i} \mathrm{Cov}(x_k) \\
&= \frac{1}{n_i}\Sigma_i.
\end{aligned}$$

Thus the estimator is unbiased, and its variance decreases at rate $1/n_i$. For simplicity and without loss of generality, we assume isotropic noise $\Sigma_i = \sigma^2 I_d$, which matches Assumption 3.1, establishing Proposition B.1. □

This derivation motivates the formal noise assumption introduced in Assumption 3.1, which models the frequency-dependent uncertainty of LM embeddings.

$$\mathbf{e}_i^{\mathrm{LM}} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{n_i}\mathbf{I}_d\right)$$

B.2 Quantifying Noise Difference Between Two Concepts (Definition 3.2)

In this subsection, we consider two independent clinical concepts and quantify the expected variability between their noisy embeddings. This result serves to define the pairwise noise scale, as introduced in Definition 3.2.

Lemma B.2 (Noise Difference Between Two Concepts)**.** *For two concepts i and j, we assume their embedding noise vectors $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_j$ are independent and satisfy*

$$\boldsymbol{\epsilon}_c \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{n_c}\mathbf{I}_d\right), \qquad c \in \{i, j\}.$$

Then the covariance of the noise difference $\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j$ is

$$\mathrm{Cov}(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\mathbf{I}_d.$$

Consequently, the expected squared magnitude of the noise difference satisfies

$$\mathbb{E}\left\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\right\|^2 = \sigma^2 d\left(\frac{1}{n_i} + \frac{1}{n_j}\right),$$

and the Root Mean Square (RMS) magnitude of the noise difference satisfies

$$\left\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\right\|_{\mathrm{RMS}} := \sqrt{\mathbb{E}\left\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\right\|^2} = \sigma\sqrt{d\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Proof. Under Assumption 3.1, the noise vectors are independent and satisfy:

$$\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \frac{\sigma^2}{n_i}\mathbf{I}_d, \qquad \mathrm{Cov}(\boldsymbol{\epsilon}_j) = \frac{\sigma^2}{n_j}\mathbf{I}_d.$$

By independence, the covariance of their noise difference is:

$$\mathrm{Cov}(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j) = \mathrm{Cov}(\boldsymbol{\epsilon}_i) + \mathrm{Cov}(\boldsymbol{\epsilon}_j) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\mathbf{I}_d.$$

Taking the trace gives the expected squared norm:

$$\mathbb{E}\left\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\right\|^2 = \mathrm{Tr}\left(\mathrm{Cov}(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j)\right) = \sigma^2 d\left(\frac{1}{n_i} + \frac{1}{n_j}\right).$$

Finally, taking the square root yields the RMS magnitude:

$$\left\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\right\|_{\mathrm{RMS}} = \sqrt{\mathbb{E}\left\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\right\|^2} = \sigma\sqrt{d\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

□

Lemma B.3 (Noise Concentration in High Dimensions)**.** *Under Assumption 3.1, for two concepts i and j, let $\boldsymbol{\epsilon}_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{n_i}\mathbf{I}_d\right)$ and $\boldsymbol{\epsilon}_j \sim \mathcal{N}\left(0, \frac{\sigma^2}{n_j}\mathbf{I}_d\right)$ be independent. Then the noise difference norm concentrates around its RMS scale:*

$$\frac{\left\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\right\|}{\sigma\sqrt{d\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \xrightarrow{p} 1 \qquad (d \to \infty),$$

and equivalently,

$$\frac{\sigma\sqrt{d\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}{\left\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\right\|} \xrightarrow{p} 1 \qquad (d \to \infty).$$

Proof. By Lemma B.2,

$$\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j \sim \mathcal{N}\Big(0,\ \sigma^2\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)\mathbf{I}_d\Big)\,,$$

and

$$\mathbb{E}\,\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|^2 = \sigma^2 d\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)\,.$$

Since the noise difference is an isotropic Gaussian vector, its squared norm divided by its expectation follows a normalized χ^2_d distribution. Therefore,

$$\frac{\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|^2}{\sigma^2 d\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)} \xrightarrow{p} 1 \qquad (d \to \infty).$$

Taking square roots yields

$$\frac{\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|}{\sigma\sqrt{d\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)}} \xrightarrow{p} 1.$$

The reciprocal convergence follows by continuous mapping. □

B.3 Noise Concentration in High Dimensions (Theorem 3.3)

Proof of Theorem 3.3. Let i and j be two low-frequency, fine-grained concepts considered in Theorem 3.3. Under Assumption 3.1, their hierarchy-agnostic LM embeddings admit the decomposition

$$\mathbf{e}_i^{\mathrm{LM}} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \qquad \mathbf{e}_j^{\mathrm{LM}} = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_j,$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 n_i^{-1}\mathbf{I}_d)$ and $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \sigma^2 n_j^{-1}\mathbf{I}_d)$ are independent.

Recall from Definition 3.2 that the pairwise signal-to-noise ratio is defined as

$$\mathrm{SNR}(i,j) = \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma\sqrt{d\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)}}.$$

By the low-SNR assumption in Theorem 3.3, this quantity satisfies $\mathrm{SNR}(i,j) \ll 1$.

By Lemma B.2, the noise difference $\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j$ is an isotropic Gaussian vector with covariance $\sigma^2\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)\mathbf{I}_d$. Lemma B.3 then implies the concentration result

$$\frac{\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|}{\sigma\sqrt{d\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)}} \xrightarrow{p} 1,$$

and equivalently,

$$\frac{\sigma\sqrt{d\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)}}{\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|} \xrightarrow{p} 1.$$

Combining the above, we write

$$\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|} = \underbrace{\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma\sqrt{d\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)}}}_{\mathrm{SNR}(i,j)} \cdot \underbrace{\frac{\sigma\sqrt{d\Big(\tfrac{1}{n_i} + \tfrac{1}{n_j}\Big)}}{\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|}}_{\xrightarrow{p} 1}.$$

The first factor converges to 0 by the low-SNR assumption, while the second factor converges in probability to 1. By Slutsky's theorem,

$$\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j\|} \xrightarrow{p} 0.$$

Therefore, for rare fine-grained concepts, the semantic separation is asymptotically dominated by estimation noise, and hierarchy-agnostic LM embeddings become geometrically inseparable, completing the proof. □

B.4 Separability Against Non-Parent Concepts under SMI (Proof of Proposition 3.6)

Recall Assumption 3.5, which states that for a leaf concept c with parent p, the true semantic representation decomposes as

$$\boldsymbol{\mu}_c = \boldsymbol{\mu}_p + \boldsymbol{\delta}_c,$$

where $\boldsymbol{\delta}_c$ is a child-specific semantic offset. Under SMI, the embedding of c aggregates representations along its hierarchical path. Since p is an ancestor of c and receives nonzero weight, the SMI semantic signal for c satisfies

$$\boldsymbol{\mu}_c^{\mathrm{SMI}} = \boldsymbol{\mu}_p + \boldsymbol{\delta}_c^{\mathrm{res}},$$

where $\boldsymbol{\delta}_c^{\mathrm{res}}$ denotes the residual child-specific component.

Let $j \neq p$ be any other concept such that $\|\boldsymbol{\mu}_p - \boldsymbol{\mu}_j\| > 0$. Then the semantic difference between c and j under SMI satisfies

$$\boldsymbol{\mu}_c^{\mathrm{SMI}} - \boldsymbol{\mu}_j = (\boldsymbol{\mu}_p - \boldsymbol{\mu}_j) + \boldsymbol{\delta}_c^{\mathrm{res}}.$$

As $n_c \to 0$, the residual term $\boldsymbol{\delta}_c^{\mathrm{res}}$ remains bounded, while the dominant semantic separation is given by $\boldsymbol{\mu}_p - \boldsymbol{\mu}_j$. Hence,

$$\|\boldsymbol{\mu}_c^{\mathrm{SMI}} - \boldsymbol{\mu}_j\| \;\longrightarrow\; \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_j\|.$$

Under Assumption 3.1, the SMI noise difference $\boldsymbol{\epsilon}_c^{\mathrm{SMI}} - \boldsymbol{\epsilon}_j$ has bounded variance because the parent-level noise does not scale with n_c. Therefore,

$$\|\boldsymbol{\epsilon}_c^{\mathrm{SMI}} - \boldsymbol{\epsilon}_j\| \;\xrightarrow{p}\; \|\boldsymbol{\epsilon}_p - \boldsymbol{\epsilon}_j\|.$$

Combining the above limits yields

$$\frac{\|\boldsymbol{\mu}_c^{\mathrm{SMI}} - \boldsymbol{\mu}_j\|}{\|\boldsymbol{\epsilon}_c^{\mathrm{SMI}} - \boldsymbol{\epsilon}_j\|} \;\xrightarrow{p}\; \frac{\|\boldsymbol{\mu}_p - \boldsymbol{\mu}_j\|}{\|\boldsymbol{\epsilon}_p - \boldsymbol{\epsilon}_j\|} \;>\; 0,$$

which establishes that SMI embeddings remain separable between c and any non-parent concept j in the long-tail regime.

□

B.5 SMI Improves SNR via Variance Reduction (Theorem 3.7 and Corollary 3.8)

For a clinical code i at any hierarchy level $\ell \in \{1, \ldots, L\}$, we model its empirical embedding as a noisy estimate of a true semantic vector $\mu_i \in \mathbb{R}^d$:

$$\mathbf{e}_i^{(\ell)} = \boldsymbol{\mu}_i^{(\ell)} + \boldsymbol{\epsilon}_i^{(\ell)}, \qquad \boldsymbol{\epsilon}_i^{(\ell)} \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{n_i^{(\ell)}}\mathbf{I}_d\right),$$

where $n_i^{(\ell)}$ is the number of samples and σ^2 is a constant noise parameter. We assume a long-tail distribution where coarser concepts are observed more frequently than fine-grained ones. Consequently, the noise variance decreases at coarser levels:

$$n_i^{(\ell)} > n_i^{(L)} \implies \frac{\sigma^2}{n_i^{(\ell)}}\mathbf{I}_d < \frac{\sigma^2}{n_i^{(L)}}\mathbf{I}_d, \qquad \forall\, \ell < L.$$

A hierarchy-agnostic LM encoder maps a fine-grained concept i (at the leaf level L) directly to an embedding, which we model as a noisy estimate of its true semantic vector:

$$\mathbf{e}_i^{\mathrm{LM}} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{n_i}\mathbf{I}_d\right).$$

Defining the semantic signal margin as $\Delta_{\mathrm{sig}}(i,j) = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$, the corresponding pairwise SNR under hierarchy-agnostic LM embeddings is

$$\mathrm{SNR}_{\mathrm{LM}}(i,j) := \frac{\Delta_{\mathrm{sig}}(i,j)}{\sqrt{\mathbb{E}\left\|\boldsymbol{\epsilon}_i^{\mathrm{LM}} - \boldsymbol{\epsilon}_j^{\mathrm{LM}}\right\|^2}} = \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{\sigma\sqrt{d\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}.$$

We first prove the result for the more general weighted SMI formulation, from which uniform SMI follows as a special case. For a code i, weighted SMI constructs a hierarchy-aware embedding by aggregating the level-specific LM embeddings of its ancestors along its hierarchical path:

$$\mathbf{e}_i^{\text{SMI-W}} = \sum_{\ell=1}^{L} w_\ell \, \mathbf{e}_{a_i^{(\ell)}}^{(\ell)}, \qquad w_\ell > 0, \quad \sum_{\ell=1}^{L} w_\ell = L,$$

where w_ℓ is the aggregation weight at level ℓ. Uniform SMI corresponds to $w_\ell = 1$ for all ℓ. For notational simplicity, we write $\mathbf{e}_i^{(\ell)}$, $\boldsymbol{\epsilon}_i^{(\ell)}$, and $n_i^{(\ell)}$ for the embedding, noise, and frequency of the level-ℓ ancestor $a_i^{(\ell)}$, respectively.

The corresponding raw noise term of the weighted sum is

$$\boldsymbol{\epsilon}_i^{\text{SMI-W}} = \sum_{\ell=1}^{L} w_\ell \, \boldsymbol{\epsilon}_i^{(\ell)}.$$

The expected squared norm of this raw noise term is

$$\mathbb{E}\left\|\boldsymbol{\epsilon}_i^{\text{SMI-W}}\right\|^2 = \sum_{\ell=1}^{L} w_\ell^2 \, \mathbb{E}\left\|\boldsymbol{\epsilon}_i^{(\ell)}\right\|^2 = d\sigma^2 \sum_{\ell=1}^{L} \frac{w_\ell^2}{n_i^{(\ell)}}.$$

For two codes i and j, the pairwise SNR of the weighted SMI embeddings is

$$\text{SNR}_{\text{SMI-W}}(i,j) = \frac{\Delta_{\text{sig}}(i,j)}{\sigma\sqrt{d\left(\sum_{\ell=1}^{L} \frac{w_\ell^2}{n_i^{(\ell)}} + \sum_{\ell=1}^{L} \frac{w_\ell^2}{n_j^{(\ell)}}\right)}}.$$

Since the semantic margin $\Delta_{\text{sig}}(i,j)$ measures the difference of true semantic representations and thus is fixed across estimators. Consequently, comparing the SNRs of hierarchy-aware (weighted) SMI embeddings and hierarchy-agnostic LM embeddings reduces entirely to comparing their noise variances.

Since pairwise SNR is scale-invariant, i.e., $\text{SNR}(\mathbf{e}) = \text{SNR}(c\,\mathbf{e})$ for any scalar $c > 0$, we can analyze a normalized version of SMI-W without changing its SNR. For weights satisfying $\sum_{\ell=1}^{L} w_\ell = L$, define $\bar{w}_\ell = w_\ell / L$, so that $\sum_{\ell=1}^{L} \bar{w}_\ell = 1$. The normalized embedding $\bar{\mathbf{e}}_i^{\text{SMI-W}} = \mathbf{e}_i^{\text{SMI-W}}/L$ has the same pairwise SNR as $\mathbf{e}_i^{\text{SMI-W}}$, so it suffices to compare its noise with that of the LM embedding.

With this normalization, we continue to use w_ℓ to denote the normalized weights, and we use the same SMI-W notation for the normalized embedding in the following variance calculation. For the normalized SMI-W embedding, the noise variance is

$$\mathbb{E}\left\|\boldsymbol{\epsilon}_i^{\text{SMI-W}}\right\|^2 = d\sigma^2 \sum_{\ell=1}^{L} \frac{w_\ell^2}{n_i^{(\ell)}}.$$

Since $w_\ell \geq 0$, expanding the square gives

$$\left(\sum_{\ell=1}^{L} w_\ell\right)^2 = \sum_{\ell=1}^{L} w_\ell^2 + 2 \sum_{1 \leq k < \ell \leq L} w_k w_\ell \geq \sum_{\ell=1}^{L} w_\ell^2.$$

Hence, under the normalized weights $\sum_{\ell=1}^{L} w_\ell = 1$,

$$\sum_{\ell=1}^{L} w_\ell^2 \leq \left(\sum_{\ell=1}^{L} w_\ell\right)^2 = 1,$$

with strict inequality when positive weights are assigned to more than one hierarchy level. Combining this with the long-tail assumption $n_i^{(\ell)} > n_i^{(L)}$ for all $\ell < L$, we obtain

$$\sum_{\ell=1}^{L} \frac{w_\ell^2}{n_i^{(\ell)}} < \frac{\sum_{\ell=1}^{L} w_\ell^2}{n_i^{(L)}} < \frac{1}{n_i^{(L)}}.$$

The strict inequalities hold because normalized SMI-W assigns positive weight to multiple hierarchy levels, including at least one ancestor level that is more frequent than the leaf concept.

so that

$$\mathbb{E}\big\|\boldsymbol{\epsilon}_i^{\text{SMI-W}}\big\|^2 = d\sigma^2 \sum_{\ell=1}^{L} \frac{w_\ell^2}{n_i^{(\ell)}} < \frac{d\sigma^2}{n_i^{(L)}} = \mathbb{E}\big\|\boldsymbol{\epsilon}_i^{\text{LM}}\big\|^2.$$

Applying this to both codes i and j gives

$$\underbrace{\left(\sum_{\ell=1}^{L} \frac{w_\ell^2}{n_i^{(\ell)}} + \sum_{\ell=1}^{L} \frac{w_\ell^2}{n_j^{(\ell)}}\right)}_{\text{normalized SMI-W noise scale}} < \underbrace{\left(\frac{1}{n_i^{(L)}} + \frac{1}{n_j^{(L)}}\right)}_{\text{LM noise scale}}.$$

This shows that the normalized SMI-W embedding has a smaller pairwise noise scale than the hierarchy-agnostic LM embedding. Since pairwise SNR is scale-invariant, the unnormalized and normalized SMI-W embeddings have the same SNR. Therefore,the unnormalized SMI-W embeddings achieve higher pairwise SNR than the hierarchy-agnostic LM embedding:

$$\text{SNR}_{\text{SMI-W}}(i,j) > \text{SNR}_{\text{LM}}(i,j).$$

Uniform SMI is a special case of SMI-W with $w_\ell = 1$ for all ℓ. Therefore, uniform SMI embeddings also achieve higher pairwise SNR than the hierarchy-agnostic LM embedding:

$$\text{SNR}_{\text{SMI}}(i,j) > \text{SNR}_{\text{LM}}(i,j).$$

B.6 Offset Consistency Analysis for Hierarchical Semantic Decomposition

Assumption 3.5 models a child clinical concept as the composition of its parent concept semantics and a child-specific semantic offset, such as a complication, manifestation, or severity. This provides an approximate representation-level decomposition of fine-grained clinical concepts. To empirically assess this assumption, We empirically assess this assumption by testing whether ICD-9 parent-child concept pairs sharing the same clinical modifier produce more consistent embedding offsets than parent-child pairs from different modifier groups. we use the clinical modifier expressed in each ICD-9 child description relative to its parent description as the child-specific semantic offset. For example, in the pair "diabetes mellitus" $\rightarrow$ "diabetes mellitus with renal manifestations," the modifier is "renal manifestations." We then test whether parent-child concept pairs sharing the same modifier produce more consistent embedding offsets than pairs from different modifier groups.

We construct the offset consistency analysis using the ICD-9 parent-child concept pairs. We define each decimal-level ICD-9 code as a child concept and use its integer-level ICD-9 code as the parent concept. Each ICD-9 code description is encoded using BioBERT with mean pooling over token embeddings. For each valid parent-child pair (p, c), we define the offset:

$$\Delta_{p,c}^{\text{LM}} = \mathbf{e}_c^{\text{LM}} - \mathbf{e}_p^{\text{LM}},$$

where $\mathbf{e}_p^{\text{LM}}$ and $\mathbf{e}_c^{\text{LM}}$ denote the BioBERT embeddings of the parent and child descriptions, respectively. This offset represents the semantic change from a coarser parent concept to a more specific child concept in the LM embedding space.

We group offsets by clinical modifiers extracted from the child description relative to the parent description. Modifier groups are extracted from the ICD-9 child description relative to its parent description using rule-based lexical patterns. We first identify explicit modifier cues, including *with*, *without*, *due to*, *secondary to*, and *associated with*; when these cues are absent, we extract residual child-description terms after removing parent-description tokens. Extracted phrases are normalized by removing generic terms and canonicalizing common variants such as *manifestation/manifestations* and *complication/complications*. We further refine broad phrases into clinically specific modifier groups, including infection status, obstruction subtype, hepatic coma status, ophthalmic manifestations, retinal disorders, postpartum complications, delivery-related complications, and loss-of-consciousness status. Vague ICD descriptions containing terms such as *unspecified*, *other*, *NEC*, *NOS*, or *not elsewhere classified* are removed, and unresolved coarse buckets such

Table 6: Strict ICD-9 offset consistency analysis. Same-modifier offsets are substantially more aligned than different-modifier clinical offsets, with a mean cosine difference of 0.575 and a one-sided p-value of 2.0×10^{-4}, supporting the hierarchical semantic decomposition assumption.

Pair type	**Mean cosine**	**Median cosine**
Same modifier	0.655	0.730
Different modifier	0.080	0.068

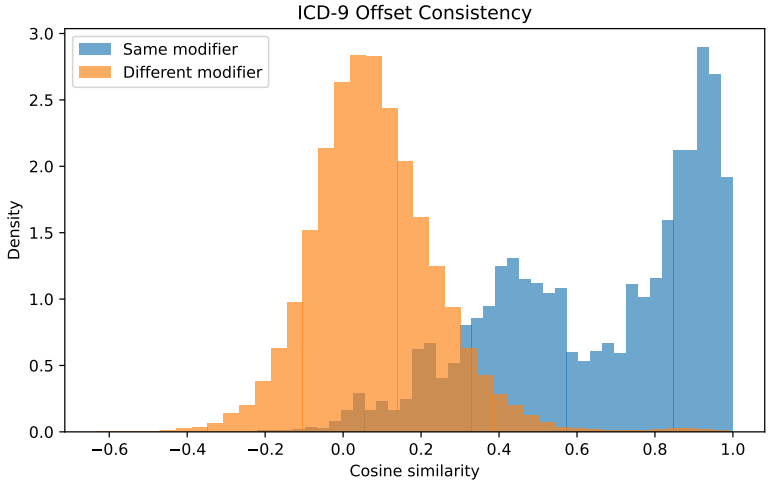

Figure 5: Distribution of cosine similarities between ICD-9 parent-child offset vectors in the strict modifier analysis. Same-modifier pairs compare offsets sharing the same extracted clinical modifier, while different-modifier pairs compare offsets from different modifier groups.The same-modifier distribution is shifted strongly to the right, indicating stronger offset consistency than random clinical offsets and supporting the hierarchical semantic decomposition assumption.

as `kidney/renal`, `circulatory/peripheral`, and `neurological` are omitted from the strict analysis because they merge multiple clinically distinct semantic changes. For each selected modifier group, we normalize the offset vectors and compute all pairwise cosine similarities among offsets within that group. These pairs serve as positive pairs, since they compare parent-child offsets sharing the same clinical modifier. For the negative pairs, we sample offset pairs from different selected modifier groups. This comparison tests whether parent-child offsets with the same clinical modifier are more directionally consistent than randomly paired offsets with different modifiers in the same embedding space.

After applying these selection criteria, the offset consistency analysis retains 418 modifier groups and 1,414 ICD-9 parent–child offsets, yielding 3,606 positive pairs with the same clinical modifier. The negative set consists of 50,000 randomly sampled offset pairs from different modifier groups. As shown in Table 6, positive pairs with the same clinical modifier have substantially higher cosine similarity than negative pairs from different modifier groups. The mean cosine similarity is 0.655 versus 0.080, with a median of 0.730 versus 0.068 and a mean difference of 0.575 ($p = 2.0 \times 10^{-4}$). Fig. 5 shows that the same-modifier distribution is substantially shifted to the right relative to the different-modifier baseline. The distribution also contains both high-similarity and moderate-similarity regions, suggesting that some modifier families induce highly stable offsets while others remain more heterogeneous. This pattern is expected because ICD-9 modifiers vary in specificity and standardization. Overall, these results show that ICD-9 parent-child concept pairs sharing the same clinical modifier produce more aligned embedding offsets than pairs from different modifier groups. This supports the hierarchical semantic decomposition in Assumption 3.5 that fine-grained clinical concepts can be represented through parent-level semantics plus a child-specific semantic offset.

C Additional Information for Medical Ontologies

C.1 Construction of Medical Ontologies

We explicitly incorporate domain knowledge to represent diagnoses, procedures, and medications using expert-defined, tree-structured medical ontologies. Each clinical concept is associated with a unique hierarchical path, capturing its semantic lineage from coarse clinical categories to fine-grained codes. This one-to-one mapping ensures that each clinical code has a distinct and interpretable position in the taxonomy. These hierarchies define the structural relationships used to construct SMIs, allowing embeddings to encode clinically meaningful relationships consistent with expert knowledge.

Diagnosis. Diagnostic concepts are organized using the CCS developed by the Agency for Healthcare Research and Quality (AHRQ). CCS provides an expert-curated framework that groups ICD-9-CM and ICD-10-CM diagnosis codes into clinically coherent categories at multiple levels of granularity. Based on this structure, we construct a five-level hierarchy comprising major categories, sub-categories, CCS codes, integer-level ICD codes, and decimal-level ICD codes. Each ICD code is uniquely assigned to a single hierarchical path, ensuring an interpretable and unambiguous disease representation.

Procedure. Procedure concepts follow the CCS procedure hierarchy, which organizes ICD-9-PCS and ICD-10-PCS codes into a multi-level taxonomy. Specifically, procedures are structured from major categories and sub-categories to CCS procedure codes, followed by three-digit ICD-PCS codes and full ICD-PCS codes. This hierarchy captures progressively finer procedural distinctions while maintaining a unique parent–child relationship at each level, resulting in a well-defined tree structure suitable for embedding learning.

Medication. Medication concepts are organized using the ATC classification system, which naturally defines a five-level hierarchy ranging from anatomical main groups to specific chemical substances. This expert-defined structure encodes pharmacological relationships among drugs and provides a clinically grounded hierarchy for representing medication concepts within the SMI framework.

Across diagnoses, procedures, and medications, each clinical code is mapped to a unique hierarchical path, yielding a unified ontology that supports effective and interpretable representation learning with knowledge-based guidance.

C.2 Statistics of Medical Ontologies

We adopt expert-defined clinical ontologies to represent diagnoses, procedures, and medications using multi-level hierarchical taxonomies. The summary statistics of clinical hierarchies are reported in Table 7.

Diagnosis: Diagnosis concepts follow the CCS hierarchy, which organizes ICD-9-CM and ICD-10-CM codes into a five-level taxonomy. 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. At the fourth level, the hierarchy contains includes 2,175 integer-level ICD codes, comprising 937 ICD-9-CM codes and 1,278 ICD-10-CM codes. At the fifth level, the hierarchy further expands to decimal-level ICD diagnostic codes, including 13,193 ICD-9-CM and 73,066 ICD-10-CM codes, thereby providing the finest granularity.

Procedure: Procedure concepts follow the CCS hierarchy, which organizes ICD-9-PCS and ICD-10-PCS codes into a five-level taxonomy. Specifically, the hierarchy comprises 16 major categories, 200 subcategories, and 224 CCS codes, covering all ICD-9-PCS and ICD-10-PCS procedure codes. At the fourth level, procedure concepts are represented by three-digit ICD-PCS codes, comprising 876 ICD-9-PCS codes and 880 ICD-10-PCS codes. The fifth level corresponds to the full ICD-PCS codes, including 3,674 ICD-9-PCS codes and 80,492 ICD-10-PCS codes.

Medication: Medication concepts follow the ATC classification system, which organizes drug codes into a five-level taxonomy. Specifically, the hierarchy comprises 14 anatomical main groups, 90 therapeutic subgroups, 248 pharmacological subgroups, 840 chemical subgroups, and 5,495 chemical substances (i.e., ATC codes).

Table 7: Hierarchical taxonomy of diagnosis, procedure, and medication concepts used in this study. Each level contains two attributes: *Name*, which specifies the semantic grouping at that level, and *# Concepts*, the number of unique codes at that level. Diagnosis categories are derived from ICD-CM codes mapped through the CCS diagnostic hierarchy; procedure categories use ICD-PCS codes organized by the CCS procedure hierarchy; and medications follow the five-level ATC ontology. For ICD-CM and ICD-PCS levels, the *# Concepts* is reported as the total number of ICD-9 and ICD-10 codes shown in parentheses, respectively.

Level	Description	Diagnosis	Procedure	Medication
Level 1	Name # Concepts	Major category 18	Major category 16	Anatomical main group 14
Level 2	Name # Concepts	Sub-category 34	Sub-category 200	Therapeutic subgroup 90
Level 3	Name # Concepts	CCS code 265	CCS code 224	Pharmacological subgroup 248
Level 4	Name # Concepts	ICD code (integer) (937, 1278)	ICD code (three-digit) (876, 880)	Chemical subgroup 840
Level 5	Name # Concepts	ICD code (13193, 73066)	ICD code (decimal) (3674, 80492)	Chemical substance 5495

D Additional Experiment Details

D.1 MIMIC Dataset Processing

MIMIC-IV is a publicly available, de-identified EHR database derived from the Beth Israel Deaconess Medical Center (BIDMC) and released through PhysioNet under the PhysioNet Credentialed Health Data License [Johnson et al., 2023]. The database contains comprehensive longitudinal hospital and intensive care data collected from emergency department and inpatient admissions between 2008 and 2022. In this study, we use the hospital module of MIMIC-IV, which provides structured information on inpatient hospitalizations, including diagnoses, procedures, medication prescriptions, and admission and discharge timestamps. To ensure patient privacy, all identifiers are replaced with de-identified integer keys, and all timestamps are shifted by a patient-specific offset. This transformation preserves within-patient temporal ordering while preventing alignment of absolute calendar time across patients. After preprocessing, our cohort comprises 223,452 unique patients and 546,028 hospital admissions. On average, each patient has 2.44 hospital visits, forming longitudinal trajectories suitable for sequential modeling.

Table 8: Basic statistics of the pre-processed MIMIC-IV EHR dataset. "#" denotes counts and "/patient" denotes per-patient averages.

Dataset	#patients	#visits	#visits/patient	#conditions/patient	#procedures/patient	#drugs/patient
MIMIC-IV	223,452	546,028	2.44	28.48	3.85	48.32

We represent the EHR data using a visit-based formulation, where each visit corresponds to one inpatient hospitalization and each patient is modeled as a temporally ordered sequence of visits. Visits are ordered chronologically by admission time, while clinical concepts within a visit are treated as unordered sets, reflecting the episodic nature of hospital care. For each visit, we extract diagnostic and procedural information from the MIMIC-IV diagnoses and procedures tables. Diagnostic concepts are represented using the original ICD-9-CM and ICD-10-CM codes, with the coding system preserved via the recorded ICD version. After preprocessing, the dataset contains 9,143 unique ICD-9 diagnosis codes and 19,440 unique ICD-10 diagnosis codes. Procedural concepts are represented using ICD-9 and ICD-10-PCS procedure codes, yielding 2,557 unique ICD-9 procedure codes and 12,354 unique ICD-10 procedure codes in total. Within each visit, diagnoses and procedures are treated as unordered sets, while visit sequences for each patient preserve temporal order across admissions. Across the

cohort, patients have on average 28.48 diagnosis codes and 3.85 procedure codes, aggregated over all visits.

Medication information is extracted from the MIMIC-IV prescriptions table, where drugs are originally recorded using National Drug Codes (NDC). Because NDC identifiers are highly granular and introduce substantial sparsity, we standardize medication concepts through a two-stage mapping process. Each NDC is first mapped to a corresponding RxNorm concept identifier (RxCUI), and each RxCUI is subsequently mapped to an Anatomical Therapeutic Chemical (ATC) classification code using a curated RxNorm–ATC reference. Unlike common practice that aggregates medications at higher ATC levels, we retain ATC level-5 codes, which correspond to specific chemical substances and preserve fine-grained pharmacological information. Not all NDC entries can be reliably mapped through this pipeline, for example due to non-drug supplies, institution-specific packaging codes, or missing references in external vocabularies. Following a strict standardization strategy consistent with prior work, unmappable NDCs are excluded from the medication representation. As a result, each visit's medication set consists only of successfully mapped ATC level-5 codes. After preprocessing, the final medication vocabulary contains 1,286 unique ATC level-5 codes. Based on the processed dataset, approximately 85% of hospital admissions contain at least one mapped ATC-5 medication, computed as the fraction of visits with a non-empty ATC-5 medication set. Across all patients, the average number of medication codes per patient is 48.32, aggregated over their longitudinal visit histories.

Based on the visit-based EHR representation, we construct training samples for multiple clinical prediction tasks using longitudinal sequences of inpatient hospitalizations. For each patient with a chronologically ordered visit sequence $\{x_1, x_2, \ldots, x_t\}$, we generate a set of samples $\{(x_1), (x_1, x_2), \ldots, (x_1, x_2, \ldots, x_t)\}$, where each sample corresponds to a prefix of consecutive visits. Each visit x_i is represented as an unordered set of diagnosis, procedure, and medication codes, while temporal dependencies are captured exclusively through the ordering of visits within the sequence. We split MIMIC-IV at the patient level, ensuring that all visits from the same patient remain in a single training, validation, or test partition. Visits are ordered chronologically, and only preceding visits are used to predict the target visit. This unified formulation enables different clinical prediction tasks to share a consistent input representation, differing only in their target definitions and the specific visit used for supervision.

We consider four clinical prediction tasks. Mortality prediction (MT) aims to predict the in-hospital mortality outcome of the subsequent visit. Formally, given an input sequence $(x_1, x_2, \ldots, x_{t-1})$, the model predicts a binary label $y[x_t] \in \{0, 1\}$, where $y[x_t] = 1$ indicates that the patient died during visit x_i; the final visit of each patient is excluded from training since no future outcome is available. Readmission prediction (RA) predicts whether the patient will be readmitted within a predefined time window of σ days. Using the same input sequence $(x_1, x_2, \ldots, x_{t-1})$, the target label is defined as $y[\tau(x_t) - \tau(x_{t-1})] \in \{0, 1\}$, where $\tau(x_i)$ denotes the admission time of visit x_i, and $y = 1$ if $\tau(x_t) - \tau(x_{t-1}) \leq \sigma$, and 0 otherwise. We set $\sigma = 15$ days. The evaluation metrics is AUPRC and AUROC. Drug recommendation is formulated as a multi-label classification task, where the input sequence $(x_1, x_2, \ldots, x_t)$ is used to predict the set of medications prescribed during visit x_t. The target is represented as a multi-hot vector over the ATC level-5 medication vocabulary, whose dimensionality equals the number of unique drug types.

D.2 eICU Dataset Processing

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 observations across hospitals using both SMI and BioBERT embedding methods. We map diagnoses and medications to 72 unique integer-level ICD-9 codes and 241 unique five-level ATC codes. We treat each patient's medical records as unordered observations without considering temporal order. We select the 10 hospitals with the largest number of patients, resulting in 2,418 patients per hospital and a total of 24,179 patients for analysis.

In eICU, approximately one-third of drug names are not recorded across the 208 hospitals (Fig. 6a). This results in an extremely sparse input and thus hampers the model training due to the lack of informative features. To address this issue, we imputed the missing drug names using the Hierarchical

Ingredient Code List (HICL) codes, a standardized coding system for healthcare products. Specifically, for entries without a drug name but with a HICL code, we retrieved the corresponding name by matching to other records that share the same HICL code and have a documented name. Using the HICL-based imputation strategy, the proportion of unrecorded drugs in our dataset was dropped to around 20% (Fig. 6b), providing denser medication inputs for downstream modeling.

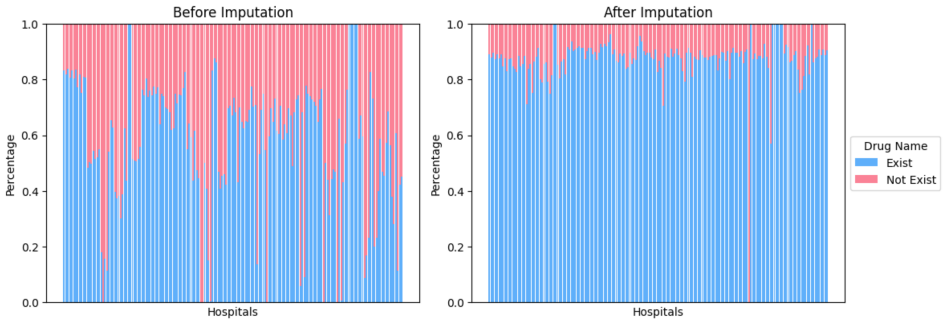

Figure 6: The drug rates before and after imputation of the eICU data. Since some drugs may lack recorded names, drug rates indicate the proportion of administered drugs with recorded names over all unique drugs.

Drug names in eICU are recorded as free-text strings with dosage information, units, and other administration-related details. Consequently, the same active ingredient can appear in many lexical variants, potentially introducing irrelevant noise for training. To address this issue, we applied a data cleaning procedure to the drug-name strings. Specifically, we retained the tokens corresponding to the core medications and removed administration-specific details such as dosages, units, and numeric values. The resulting drug texts more closely reflect the medication's active-ingredient terms, supporting more stable and transferable medication representations.

We also found that the same medication can appear under different names in the eICU dataset (e.g., trade/brand names vs. ingredient names), creating multiple lexical variants of the same drug entity. This variability inflates the effective vocabulary and introduces spurious distinctions that reduce model robustness. To mitigate this issue, we harmonized eICU drug strings by mapping them to standardized medication concepts using the WHO Anatomical Therapeutic Chemical (ATC) classification system. Specifically, we first collected ATC classes from the WHO Collaborating Centre for Drug Statistics Methodology website [Kury, 2024]. We then leveraged the five-level ATC hierarchy and extracted the 5th-level ATC substance codes as well as their descriptions. For each eICU canonical drug name, we computed BioWordVec embeddings using for both the eICU string and each reference ATC substance name [Zhang et al., 2019]. We then measured the cosine similarity between these two word embeddings. After comparing with all reference drug names, the reference ATC substance name with the highest cosine similarity is selected. We also assigned the corresponding ATC code to each medication record. This procedure effectively matches drugs with similar ATC codes, improves the consistency of medication concepts across records, and supports standardized analyses across ATC levels.

D.3 Prompt for Biomedical LMs

Specifically, we use a frozen Biomedical LM 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 9 presents the BioBERT input prompt, in which the text passed to the LM encoder explicitly encodes the hierarchical context of a clinical concept.

D.4 Baseline Details

We compare SMI against representative baseline methods for learning medical concept embeddings. These baselines differ in how they incorporate textual information, ontology structure, and embedding

Table 9: Level-specific textual descriptions used by the biomedical LM. The example includes an ICD-10 diagnosis (*I15.1 Hypertension secondary to other renal disorders*), an ICD-10 procedure (*79658 Magnetic Resonance Imaging (MRI) of Lumbar Spine*), and an ATC medication (*A10AE04 Insulin glargine*). For each hierarchy level, we encode only the textual description of the code at that level, rather than concatenating descriptions along the full ontology path.

Level	Diagnosis	Procedure	Medication
Level 1	Diseases of circulatory system	Miscellaneous diagnostic and therapeutic procedures	Alimentary tract and metabolism
Level 2	Hypertension	Magnetic resonance imaging	Drugs used in diabetes
Level 3	Hypertension with complications and secondary hypertension	Magnetic resonance imaging	Insulins and analogues
Level 4	Secondary hypertension	Imaging, Axial Skel, Exc Skull Facial, MRI Imaging, Axial Skeleton, Except Skull and Facial Bones, Magnetic Resonance Imaging (MRI)	Insulins and analogues for injection, long-acting
Level 5	Hypertension secondary to other renal disorders	Magnetic Resonance Imaging (MRI) of Lumbar Spine	Insulin glargine

geometry, providing complementary points of comparison for evaluating hierarchy awareness and robustness under long-tailed concept distributions. We include both concept-level and patient-level representation baselines, as several prior clinical models derive concept embeddings implicitly through task-driven training.

BioBERT. BioBERT encodes clinical concepts by embedding their textual descriptions using a pretrained biomedical language model [Lee et al., 2020], yielding hierarchy-agnostic Euclidean representations. While effective for frequent concepts, BioBERT does not explicitly model hierarchical relationships among medical codes and tends to degrade for sparse, fine-grained concepts.

SapBERT. SapBERT learns biomedical concept representations by aligning synonymous terms in UMLS using contrastive learning over pretrained biomedical language models [Liu et al., 2021]. By enforcing proximity among synonym sets, SapBERT improves semantic consistency and entity normalization performance. However, its supervision is limited to local synonym relations and does not explicitly model hierarchical or multi-level ontology structure, making it insufficient for preserving coarse-to-fine semantics across medical taxonomies.

G-BERT. G-BERT incorporates medical ontology information into a Transformer model by augmenting patient records with graph-based contextual representations and pretraining objectives [Shang et al., 2019]. While effective for medication recommendation, G-BERT focuses on patient-level sequence modeling and does not produce standalone hierarchy-aware concept embeddings.

HiTs. HiTs treats pretrained language models as implicit hierarchy encoders by supervising them to produce representations that respect hierarchical relations among concepts [He et al., 2024]. While effective at aligning LM embeddings with taxonomic structure through structured objectives, HiTs relies on learning hierarchy consistency during training and does not explicitly aggregate multi-level ancestor context to form a single coarse-to-fine concept embedding.

OnT. OnT treats pretrained language models as ontology encoders by fine-tuning them under geometric constraints in hyperbolic space to preserve logical relations defined in OWL ontologies [Yang et al., 2025]. By jointly incorporating textual labels and formal ontology axioms, OnT achieves strong hierarchy consistency and supports ontology-level reasoning tasks. In contrast to SMI, OnT focuses on enforcing global logical structure through geometric modeling, rather than aggregating multi-level semantic signals along ontology paths for downstream clinical representation learning.

GRAM. GRAM learns medical concept representations by recursively aggregating ancestor information in a medical ontology using an attention mechanism [Choi et al., 2017]. By leveraging hierarchical structure, GRAM improves embeddings for rare concepts, but it is trained in a supervised, task-specific manner rather than producing general-purpose concept embeddings.

cui2vec. cui2vec learns distributed embeddings of medical concepts by applying word embedding methods to large-scale clinical notes, claims data, and biomedical literature [Beam et al., 2020]. These embeddings capture statistical co-occurrence patterns but do not explicitly encode hierarchical or relational structure from medical ontologies.

RotatE. RotatE embeds entities in complex space and models relations as rotations, enabling it to capture symmetry, antisymmetry, and inversion patterns in knowledge graphs [Sun et al., 2019]. While more expressive than TransE, RotatE does not explicitly encode hierarchical depth or ontology-defined parent–child semantics.

BEHRT. BEHRT is a Transformer-based model pretrained on longitudinal EHR sequences using diagnosis codes, age, and temporal embeddings to learn patient representations [Li et al., 2020]. While effective for clinical risk prediction, BEHRT is designed for patient-level modeling rather than reusable medical concept embeddings.

Med-BERT. Med-BERT pretrains a BERT-style model on structured EHR data using masked code prediction and visit-level objectives to capture medical context [Rasmy et al., 2021]. Although Med-BERT improves downstream clinical tasks, it focuses on contextual patient representations rather than explicit hierarchy-aware concept encoding.

ClinicalBERT. ClinicalBERT adapts BERT to the clinical domain by pretraining on MIMIC clinical notes to learn contextualized representations of medical language [Alsentzer et al., 2019]. These embeddings capture rich clinical semantics from text but do not explicitly encode medical ontologies or hierarchical relations among concepts.

BERGAMOT. BERGAMOT learns biomedical entity representations by combining LM-based concept encoding with graph-augmented representation learning [Sakhovskiy et al., 2024]. It jointly leverages textual concept information and biomedical graph structure to improve entity representations. This requires additional graph-based representation learning and does not explicitly construct a coarse-to-fine concept representation by aggregating level-specific embeddings along an expert-defined ontology path.

BoxLM. BoxLM is a recent ontology-aware clinical prediction model that combines LM-based semantic representations with box embeddings to capture hierarchical structure for diagnosis prediction [Tan et al., 2025]. It represents medical concepts using box geometry and incorporates ontology-driven structure into a task-specific prediction framework. BoxLM is optimized for supervised prediction and does not provide a training-free, task-agnostic concept representation constructed from frozen LM embeddings along ontology paths.

D.5 Computational resources

All experiments were conducted on an NVIDIA RTX PRO 6000 GPU with 96GB memory. SMI is training-free and only requires a one-time computation of frozen LM embeddings, after which the embeddings are reused in downstream tasks. Therefore, SMI introduces minimal additional hardware or computational overhead beyond the baseline embedding extraction and downstream evaluation.

E Additional Results

E.1 Long Tail Distribution of Clinical Concepts

Clinical concepts in EHRs typically exhibit highly skewed, long-tailed distributions: a small subset of clinical codes appear frequently, while most codes are observed with low frequencies. To quantify this pattern in MIMIC-IV dataset, we compute patient-level frequencies for three coding systems:

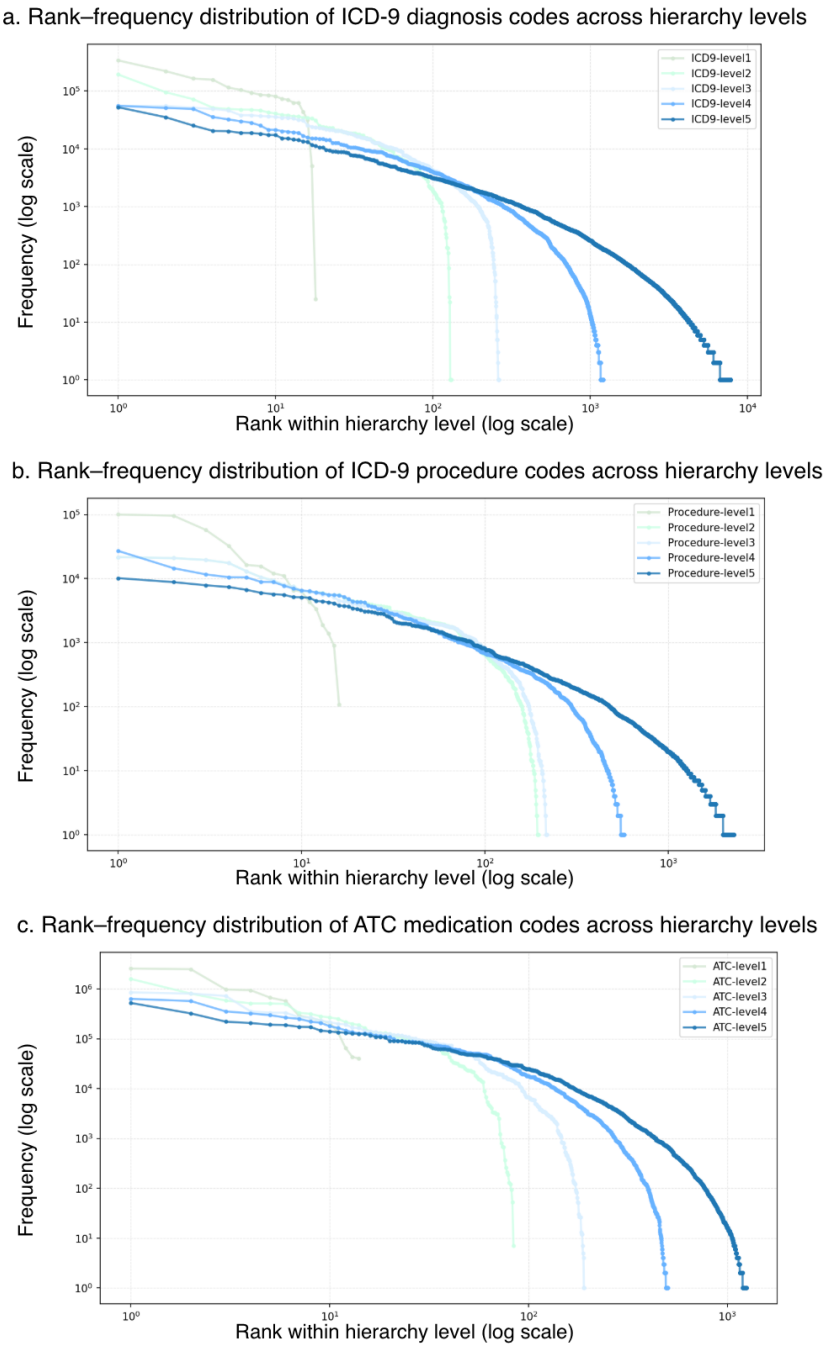

Figure 7: Rank–frequency distributions (log–log scale) of clinical concepts in MIMIC-IV across five hierarchy levels: (a) ICD-9 diagnosis codes, (b) ICD-9 procedure codes, and (c) ATC medication codes. For each hierarchy level, concepts are ranked by patient frequency within that level.

ICD-9 diagnoses, ICD-9 procedures, and ATC medications. For each hierarchy, we measure the number of unique patients associated with each concept at all five levels, rank concepts by frequency, and plot the resulting rank–frequency curves on log–log scales.

Fig. 7 summarizes the rank–frequency distributions across hierarchy levels for ICD-9 diagnoses, ICD-9 procedures, ATC medications. A consistent long-tail pattern emerges in all three hierarchies: coarse-grained levels include fewer concepts with higher patient frequencies, whereas fine-grained levels contain many more concepts whose frequencies drop sharply, yielding a large fraction of rare codes. This significant sparsity at fine granularity makes it difficult for hierarchy-agnostic LM embeddings to learn meaningful representations from limited observations. These results motivate hierarchy-aware SMI approach that integrates expert-defined hierarchical semantics to improve representations for infrequent, fine-grained clinical concepts.

E.2 Details about Clinical Relation Prediction

We design our multi-hop prediction task within the ICD and ATC hierarchies following the setup of prior work [Ganea et al., 2018]. The goal is to evaluate whether the model can infer indirect, multi-hop subsumption relations from direct, one-hop subsumptions. For each positive pair where one concept subsumes another, we sample 10 negative parent candidates that are not connected to the child concept via either direct or indirect subsumption, leading to 10 training triplets per positive example We consider two negative-sampling strategies: (1) hard negatives, chosen from sibling concepts (i.e., concepts sharing the same parent); (2) random negatives, uniformly from the remaining concepts. When a concept has fewer than 10 siblings, we fill in with random negatives to maintain 1-to-10 ratio of positives to negatives.

For cross-ontology relation prediction [Zou et al., 2022], we evaluate a one-hop relation prediction using ICD→ATC-10 pair. Given an ATC medication code, we predict whether a candidate ICD-10 diagnosis code is linked to it. Since each ATC maps to a single ICD-10 label in our dataset, we restrict evaluation to 1-hop prediction and sample one negative ICD-10 code per positive pair.

For both multi-hop prediction within the ICD and ATC hierarchies and the cross-ontology one-hop ICD→ATC task, we report Precision, Recall, and F-score as the evaluation metrics. We do not use accuracy metric because class imbalance can inflate it; as the large number of negatives can cause misleadingly high scores.

E.3 Details about Clustering Analysis

We evaluate how well the embeddings group clinical codes into highest-level clinical group (i.e., major categories for ICD diagnostic and procedure codes, and Anatomical main groups for ATC medication codes). To do so, we apply K-means clustering to both BioBERT and SMI embeddings, setting the number of clusters K to match the number of ground-truth categories. We evaluate clustering quality using three groups of metrics: (1) Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Purity measure cluster-label alignment with ICD major categories; (2) Per-category Recall and Precision measure how coherently each major category is grouped and how pure its dominant cluster is; (3) Silhouette and Intra/Inter Ratio measure whether embeddings from the same major category are compact and separated from other categories. For all three metric groups, we report the aggregate clustering results in Table 10. For per-category Recall and Precision across the 18 ICD major categories, we report detailed results in Table 11.

(1) Cluster-label alignment metrics. We first report global clustering metrics that evaluate how well the unsupervised clusters align with ICD major categories. Specifically, we use NMI, ARI, and cluster Purity. NMI measures the mutual information agreement between cluster assignments and major-category labels. ARI measures pairwise agreement between the clustering result and the ground-truth labels after correcting for chance. Purity measures the fraction of samples in each cluster that belong to its dominant major category. Higher values indicate stronger ontology-consistent clustering.

(2) Aggregate and per-category Recall and Precision. We next report Recall and Precision to evaluate category-level clustering quality. For each ICD major category, we compute per-category Recall and Precision, and then report their macro-averages as aggregate metrics in Table 10. Detailed

Table 10: Aggregate clustering quality of ICD diagnosis embeddings by major disease category. NMI, ARI, Purity, Recall, Precision, and Silhouette are higher better, while Intra/Inter Ratio is lower better. SMI achieves stronger cluster-label alignment and better geometric separation than BioBERT, indicating more ontology-consistent diagnosis embeddings.

Method	**NMI** ↑	**ARI** ↑	**Purity** ↑	**Recall** ↑	**Precision** ↑	**Sil.** ↑	**Intra/Inter** ↓
SMI	**0.77**	**0.56**	**0.78**	**0.80**	**0.61**	**0.52**	**0.29**
BioBERT	0.39	0.25	0.51	0.41	0.35	0.04	0.82

per-category results are reported in Table 11. Let c index the resulting clusters, and let $c^*(m)$ denote the dominant cluster for category m, defined as the cluster containing the largest number of codes from m.

Per-category Recall measures how well a diagnostic major m is captured by a single cluster. It is , it defined as the proportion of ICD codes from major m that are assigned to its dominant cluster $c^*(m)$:

$$\text{recall}(m) = P(\text{cluster} = c^*(m) \mid \text{major} = m)$$

Higher Recall indicates that codes from the same major category are concentrated in one coherent cluster, while lower Recall indicates that the category is scattered across multiple clusters.

Per-category Precision measures the purity of the dominant cluster corresponding to a diagnostic major m. It is defined as the proportion of ICD codes within the dominant cluster $c^*(m)$ that belong to major m:

$$\text{precision}(m) \;=\; P\big(\text{major} = m \mid \text{cluster} = c^*(m)\big)$$

Higher Precision indicates that the dominant cluster contains few codes from other major categories, while lower Precision indicates stronger mixing across categories.

(3) Embedding-space separation metrics. We also report Silhouette score and Intra/Inter Ratio as global embedding-space separation metrics to assess whether ICD major categories are geometrically separated. Silhouette score measures whether each code is closer to codes in its own cluster than to codes in other clusters, where higher values indicate better separation. Intra/Inter Ratio is defined as the average within-major-category distance divided by the average between-major-category distance. Lower values indicate more compact embeddings within the same major category and stronger separation across different major categories.

Table 10 shows that SMI improves cluster-label alignment over BioBERT, with higher NMI (0.77 vs. 0.39), ARI (0.56 vs. 0.25), and Purity (0.78 vs. 0.51). This indicates that the clusters induced by SMI embeddings are more consistent with ICD major disease categories. SMI also improves per-category clustering quality, achieving higher average Recall (0.80 vs. 0.41) and Precision (0.61 vs. 0.35), showing that codes from the same major category are grouped more coherently and that the dominant clusters are less mixed with codes from other categories. Table 11 further shows that this improvement holds across most ICD major categories. Finally, SMI improves embedding-space separation, with a higher Silhouette score (0.52 vs. 0.04) and a lower Intra/Inter Ratio (0.29 vs. 0.82), indicating that codes within the same major category are more compact while different major categories are better separated. Overall, these results support that SMI produces diagnosis embeddings that are more aligned with the ICD ontology than BioBERT embeddings.

Across most majors, SMI achieves high recall, suggesting that codes within each major are concentrated in a dominant cluster rather than scattered across multiple clusters. In particular, SMI reaches recall above 80% for most ICD major categories, including Endocrine & metabolic, Blood diseases, Mental disorders, Nervous system & sense organs, Circulatory system, Respiratory system, Digestive system, Genitourinary system, Skin, Musculoskeletal, Congenital anomalies, and External causes. Moreover, SMI also achieves high precision for many majors, including 100% precision for Infectious & parasitic, Neoplasms, Mental disorders, Musculoskeletal, Congenital anomalies, Symptoms & ill-defined, and External causes. This indicates that the dominant clusters for these majors are highly pure and contain few codes from other majors.

In contrast, BioBERT exhibits lower precision across most majors, implying stronger mixing of majors within the dominant clusters even when recall is moderate. We also observe a few challenging categories where precision remains low despite high recall, such as Blood diseases, Skin, and

Table 11: Per-major clustering recall and precision by ICD major disease category. Recall measures whether codes from the same major category are concentrated in a dominant cluster, while precision measures the purity of that dominant cluster. SMI achieves higher recall and precision than BioBERT for most ICD major categories, indicating more coherent and less mixed ontology-aligned clusters.

Major	Major Category	Recall (%)		Precision (%)	
		SMI	BioBERT	SMI	BioBERT
1	Infectious & parasitic	**73.3**	33.1	**100.0**	45.4
2	Neoplasms	**62.8**	53.7	**100.0**	91.5
3	Endocrine & metabolic	**85.3**	19.8	**21.8**	8.6
4	Blood diseases	**91.7**	41.7	**10.6**	4.8
5	Mental disorders	**89.9**	58.1	**100.0**	58.7
6	Nervous system & sense organs	**89.5**	28.4	**78.2**	36.5
7	Circulatory system	**85.6**	28.6	**41.7**	17.1
8	Respiratory system	**80.9**	25.8	**24.5**	11.3
9	Digestive system	**91.7**	29.8	**53.2**	23.3
10	Genitourinary system	**85.2**	20.1	**33.2**	12.6
11	Pregnancy & childbirth	51.2	**90.2**	**85.3**	79.4
12	Skin	**87.6**	17.7	**13.4**	2.9
13	Musculoskeletal	**96.3**	55.4	**100.0**	63.4
14	Congenital anomalies	**93.5**	44.6	**100.0**	24.4
15	Perinatal conditions	**63.0**	53.5	**18.1**	15.9
16	Symptoms & ill-defined	**45.1**	31.2	**100.0**	89.5
17	Injury & poisoning	**77.1**	31.3	**18.2**	5.4
18	External causes (E-codes)	**82.8**	78.5	**100.0**	42.6

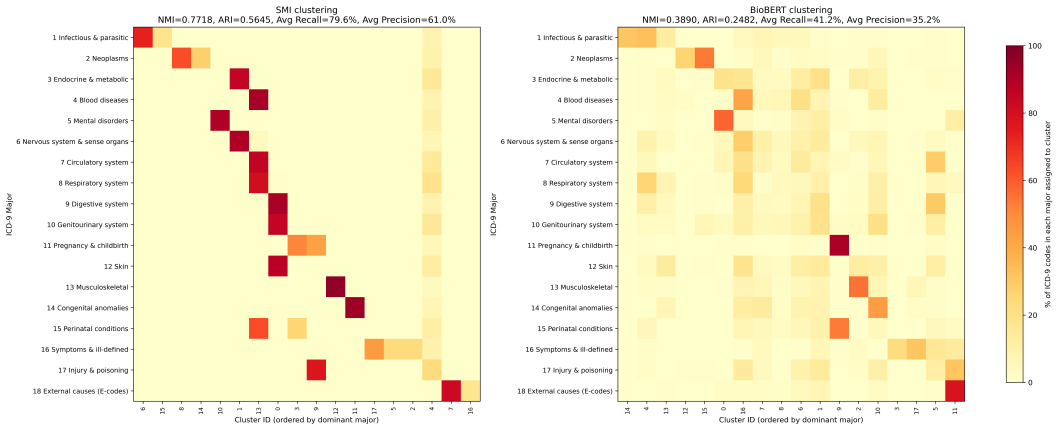

Figure 8: Cluster assignment heatmaps for ICD diagnosis embeddings. Rows denote ICD major disease categories and columns denote K-means clusters ordered by their dominant major category. Each cell shows the percentage of codes from a major category assigned to a cluster. Compared with BioBERT, SMI exhibits a stronger diagonal pattern, indicating better alignment between embedding structure and ICD major categories.

Injury & poisoning. This suggests that K-means tends to group these categories with clinically or semantically related categories, leading to mixed dominant clusters. Despite this difficulty, SMI still achieves higher recall and precision than BioBERT for these categories. Overall, the results indicate that hierarchy-aware concept representations form clusters that are more consistent with medical ontologies.

E.4 Details about Rare Code Prediction

To benchmark drug recommendation on rare ATC codes, we construct a new MIMIC-IV subset. We select five-level ATC codes with 10–20 occurrences, yielding 98 unique long-tailed medications,

and restrict drug recommendation to this label space. Because rare targets are extremely sparse, the previous threshold-based metrics (AUPRC, F-score, and Jaccard) are less informative; instead, we evaluate retrieval quality using Recall@15, i.e., whether the true medication labels appear within the top-15 prediction set.

E.5 Details about Distribution shift

We assess cross-hospital distribution shift following prior work by learning a hospital-specific density model on patient embeddings and measuring the log-likelihood gap between in-domain samples and samples from other hospitals [Zhu et al., 2025]. We trained 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 encoded diagnoses and medications using either SMI or BioBERT, and then trained a VAE density estimator $p_k(x_k)$ on the embeddings x_k from hospital k. We quantified the distribution shift from source hospital k to target hospital t by computing their difference of log-likelihoods:

$$\Delta = \mathbb{E}_{x_k}\left[\log p_k(x_k)\right] - \mathbb{E}_{x_t}\left[\log p_k(x_t)\right] \tag{6}$$

where $\log p_k(x_k)$ and $\log p_k(x_t)$ are the log-likelihoods of in-domain samples x_k and out-of-domain samples x_t evaluated under the VAE trained on k-th hospital, respectively. A large Δ implies greater cross-hospital shift between source and target distributions. We approximate the log-likelihood $\log p_k(x)$ using an Importance Weighted Autoencoder (IWAE) objective by drawing $K = 64$ importance samples $z_i \sim q(z \mid x)$ [Burda et al., 2015]:

$$\log p_k(x) \approx \log\left(\frac{1}{K}\sum_{i=1}^{K}\frac{p(x, z^{(i)})}{q(z^{(i)} \mid x)}\right) \tag{7}$$

Fig. 9 visualizes cross-hospital distribution shift as the absolute log-likelihood difference ($|\Delta|$) between each source-target hospital pair using a hospital specific density model trained on the source. SMI yields consistently lower $|\Delta|$ across source–target hospital pairs than BioBERT, indicating better cross-site stability and reduced domain shift. This improvement likely comes from injecting expert-curated medical hierarchies, which place semantically related concepts closer in the embedding space. In contrast, BioBERT relies mainly on textual similarity and does not explicitly encode clinical ontology relationships, leading to larger cross-hospital variability. Overall, SMI effectively mitigates cross-hospital distribution shift and is more suitable for multi-institutional settings with heterogeneous data.

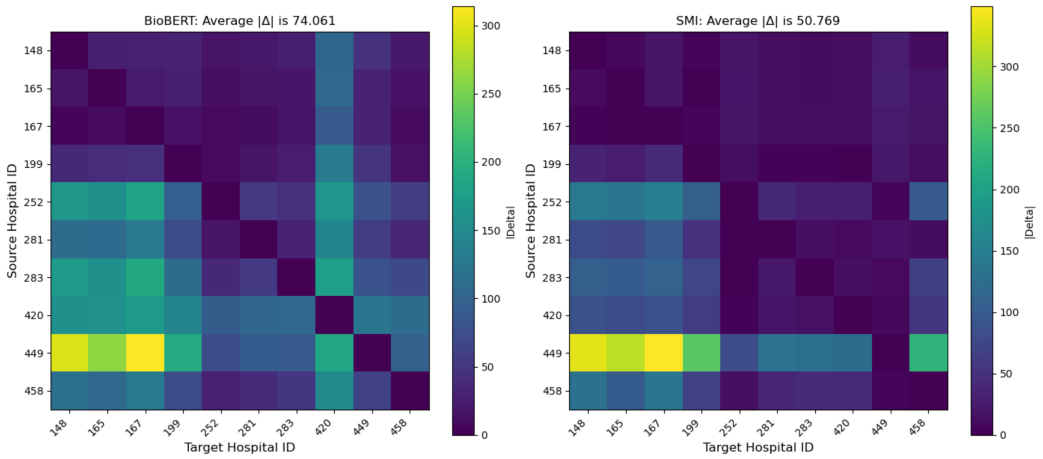

Figure 9: Cross-hospital distribution shift analysis based on average absolute log-likelihood differences ($|\Delta|$) between source and target hospitals. Each heatmap illustrates the distribution shift of patient embedding using SMI (left) and BioBERT (right), both evaluated with a VAE-based density model. SMI embeddings exhibit smaller distribution shifts (average $|\Delta| = 50.77$) compared to BioBERT ($|\Delta| = 74.06$), indicating that SMI is more robust to distribution shift.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 abstract and introduction state the main scope and contributions of SMI, including training-free hierarchy-aware concept representation, theoretical SNR analysis, experimental contributions including relation prediction, downstream clinical prediction, and cross-hospital robustness. These claims are supported by the methodology, theory, and experiments reported in the paper.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 conclusion discusses the scope and limitations of SMI, including its focus on expert-curated coding systems such as ICD and ATC, the need for additional ontology validation and code mapping for emerging disease classifications, newly introduced drugs, and local coding standards.

Guidelines:

- The answer [N/A] 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: [Yes]

   Justification: The paper states the assumptions and formal theoretical results in Section 3.3. Complete proofs and supporting derivations are provided in Appendix B.

   Guidelines:

   - The answer [N/A] 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 describes the datasets, ontology construction, data splits, tasks, metrics, baselines, and evaluation protocols in the main text and appendix. These details provide sufficient information to reproduce the main empirical results.

Guidelines:

- The answer [N/A] 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: [No]

Justification: The paper uses public but credentialed-access clinical datasets, including MIMIC-IV and eICU, whose redistribution is restricted by data-use agreements. We provide implementation details in the appendix, but the current submission does not include an anonymized code release.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the clinical prediction tasks, evaluation metrics, data splits, baselines, and implementation details in Section 4.1, 4.4 and 4.5, with additional baseline and experimental details in Appendices D.4, E.2, and E.5.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: The clinical prediction results report mean and standard error from bootstrap, while the deterministic relation prediction experiments use threshold selection via grid search and are reported without variance. The paper reports statistical significance tests for drug recommendation, rare drug prediction, and cross-hospital distribution shift.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: All experiments were conducted on an NVIDIA RTX PRO 6000 GPU with 96GB memory. SMI is training-free and only requires a one-time computation of frozen LM embeddings, after which the embeddings are reused in downstream tasks. Therefore, SMI introduces minimal additional hardware or computational overhead.

   Guidelines:

   - The answer [N/A] 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: The study uses de-identified, credentialed-access clinical datasets and does not involve intervention on patients, crowdsourcing, or collection of new human-subject data. We found no deviation from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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 discusses positive impacts through improved representation of long-tailed clinical concepts, better clinical prediction, and robustness to cross-site distribution shift. It also discusses deployment-related risks, including dependence on correct ontology selection, local code mapping, and clinician validation before real-world use. The study uses de-identified clinical datasets and does not release patient-level data.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk AI model or dataset. The clinical datasets used in the study are controlled-access and governed by their original data-use agreements. item[] Guidelines:

    - The answer [N/A] 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: The paper cites the original sources for existing datasets (i.e., MIMIC-IV, eICU), medical ontologies (i.e., ICD, CCS), language models (i.e., BioBERT), and baseline methods. Access to MIMIC-IV and eICU follows their official credentialed-use requirements.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: We plan to release the precomputed SMI concept embeddings and the code for constructing them from public medical ontologies and frozen biomedical LM encoders. We will provide documentation describing the embedding construction procedure, required ontology resources, and usage instructions for downstream tasks.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, participant recruitment, or direct interaction with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The study uses de-identified, controlled-access EHR datasets and does not collect new human-subject data or interact with participants. Dataset access is governed by the original data-use agreements.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core method uses a frozen biomedical language model encoder, BioBERT, to obtain level-specific clinical code embeddings. The paper describes this usage in the methodology.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.