OC-Distill: Ontology-aware Contrastive Learning with Cross-Modal Distillation for ICU Risk Prediction

Zhongyuan Liang ZHONGYUAN_LIANG@BERKELEY.EDU
Computational Precision Health
UC Berkeley, UCSF

Junhyung Jo JUNBRO.JO@SAMSUNG.COM
Samsung Advanced Institute of Technology (SAIT)

Hyang-Jung Lee HJEVA.LEE@SAMSUNG.COM
Samsung Advanced Institute of Technology (SAIT)

Sang Kyu Kim SANGQ.KIM@SAMSUNG.COM
Samsung Advanced Institute of Technology (SAIT)

Irene Y. Chen IYCHEN@BERKELEY.EDU
Computational Precision Health
UC Berkeley, UCSF

Abstract

Early prediction of severe clinical deterioration and remaining length of stay can enable timely intervention and better resource allocation in high-acuity settings such as the ICU. This has driven the development of machine learning models that leverage continuous streams of vital signs and other physiological signals for real-time risk prediction. Despite their promise, existing methods have important limitations. Contrastive pretraining treats all patients as equally strong negatives, failing to capture clinically meaningful similarity between patients with related diagnoses. Meanwhile, downstream fine-tuning typically ignores complementary modalities such as clinical notes, which provide rich contextual information unavailable in physiological signals alone. To address these challenges, we propose OC-Distill, a two-stage framework that leverages multimodal supervision during training while requiring only vital signs at inference. In the first stage, we introduce an ontology-aware contrastive objective that exploits the ICD hierarchy to quantify patient similarity and learn clinically grounded representations. In the second stage, we fine-tune the pretrained encoder via cross-modal knowledge distillation, transferring complementary information from clinical notes into the model. Across multiple ICU prediction tasks on MIMIC, OC-Distill demonstrates improved label efficiency and achieves state-of-the-art performance among methods that use only vital signs at inference.
Code: https://github.com/the-chen-lab/OC-Distill

1. Introduction

Early prediction of clinical deterioration in the intensive care unit (ICU) is critical for timely intervention and effective resource allocation. The availability of continuous physiological

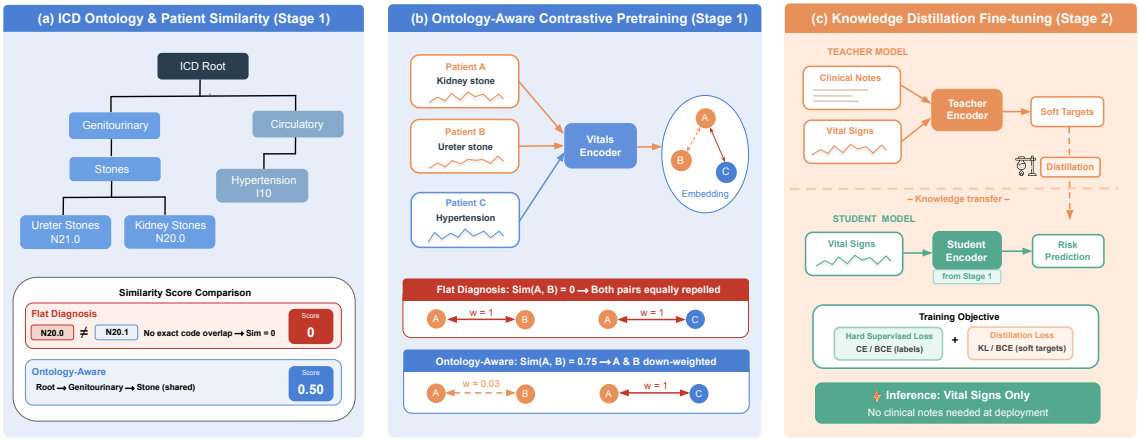

Figure 1: Overview of our two-stage framework. (a) We compute patient similarity from the ICD diagnosis hierarchy, providing more informative scores than flat diagnosis matching. (b) Ontology-aware contrastive pretraining down-weights clinically similar negative pairs while strongly repelling dissimilar pairs. (c) Knowledge distillation finetuning transfers representations from a multimodal teacher to a vitals-only student, enabling inference without clinical notes at deployment.

monitoring has enabled the development of machine learning (ML) models that can identify patients at elevated risk of adverse outcomes by processing streams of vital signs and other bedside time-series data (Suresh et al., 2017; Purushotham et al., 2018; Wang et al., 2020; Hu et al., 2026). Despite substantial progress, current approaches face two key limitations that constrain their effectiveness in real-world ICU settings.

First, contrastive learning (CL) has shown promise for pretraining physiological signals, yet existing methods treat all patients as equally strong negatives during training (Chen et al., 2020; Raghu et al., 2023; King et al., 2024), ignoring that patients often share related disease profiles. Recent work has attempted to address this by computing the percentage of exact ICD code matches between patients as a proxy for clinical similarity (Ding et al., 2024), but this flat matching approach overlooks the hierarchical structure of diagnosis taxonomies. Because the ICD taxonomy contains thousands of specific codes, exact matches are sparse even among clinically related patients. For example, a patient with ureter stones and a patient with kidney stones share substantial pathophysiological commonalities encoded in the ICD hierarchy, yet flat matching assigns them zero similarity. Consequently, pretraining strategies that ignore these hierarchical relationships may learn representations that fail to reflect clinical similarity.

Second, training on physiological time series alone ignores other rich modalities in the electronic health record (EHR). In particular, clinical notes contain clinical reasoning and contextual information, such as patient history and behavioral observations, that are not captured in physiological signals alone (Seinen et al., 2025; Liang et al., 2025b; Du et al., 2026). While recent work explores multimodal fusion approaches (Shukla and Marlin, 2020; Wang et al., 2022; Lyu et al., 2023; Battula et al., 2024; Meng et al., 2026), these methods

require all modalities as inputs at inference time—an assumption that often fails in practice. For example, clinical notes are time-consuming to generate, whereas ICU workflows demand rapid, continuous monitoring. Furthermore, different modalities are often stored in disparate systems, making real-world deployment of such approaches impractical.

To address these limitations, we introduce OC-Distill (**O**ntology-aware **C**ontrastive learning with cross-modal **Distill**ation), a two-stage framework for ICU risk prediction that leverages multimodal supervision during training while requiring only vital signs at inference (Fig 1). In the first stage, we develop an ontology-aware CL approach that incorporates patient similarity from the ICD diagnosis hierarchy. Rather than applying a uniform contrastive penalty to all negative pairs, we down-weight clinically similar patients during pretraining, allowing the model to learn representations that preserve the structure of disease relationships. In the second stage, we finetune the pretrained encoder via cross-modal knowledge distillation (Hinton et al., 2015; Lan and Tian, 2025; Duan, 2026; Lan et al., 2026b; Ding et al., 2026), where a teacher model trained on both vital signs and clinical notes transfers its representations to a student model initialized from the pretrained encoder that operates exclusively on vitals. This enables the student to benefit from clinical context during training without requiring multimodal inputs at deployment.

We evaluate our approach on MIMIC (Harutyunyan et al., 2019; Johnson et al., 2023) across multiple ICU prediction tasks, label fractions, and observation horizons. Across these settings, incorporating similarity from the ICD ontology into contrastive pretraining yields more clinically informed representations, demonstrating strong label efficiency in both linear evaluation and end-to-end fine-tuning. Furthermore, finetuning with knowledge distillation from clinical notes provides additional gains, and the full OC-Distill pipeline achieves state-of-the-art performance among methods that operate on vital signs alone at inference time. In summary, we make the following contributions:

1. An ontology-aware CL method that integrates patient similarity derived from the ICD ontology, yielding more clinically grounded and label-efficient representations.
2. A multimodal distillation approach that distills knowledge from clinical notes into a vitals model, improving downstream performance without requiring multimodal inputs at inference.
3. A comprehensive empirical evaluation on the MIMIC benchmarks, demonstrating state-of-the-art performance across tasks, label fractions, and observation horizons.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work offers several insights for machine learning in healthcare beyond the specific task of ICU risk prediction. First, our results show that incorporating domain-specific clinical structure can meaningfully improve machine learning performance. In particular, diagnostic ontologies such as ICD provide a readily available and clinically meaningful source of supervision. When used in contrastive pretraining, this structure can substantially reduce the amount of labeled data needed to achieve strong downstream performance. Second, we identify a fundamental tension between current multimodal fusion approaches and the constraints of real-world clinical deployment. In practice, data modalities are distributed

across separate systems, and some modalities, such as clinical notes, are time-consuming to generate. At the same time, ICU workflows require rapid and continuous monitoring. Our framework addresses this mismatch through knowledge distillation, allowing models to benefit from rich multimodal supervision during training while remaining practical and efficient at deployment. Overall, this work presents a generalizable approach for leveraging clinical domain structure in the learning pipeline while respecting the practical constraints of real-world deployment.

2. Related Work

2.1. Contrastive Pretraining for Clinical Time Series

CL is widely used for pretraining representations from unlabeled data due to its simple objective and strong empirical performance (Jaiswal et al., 2020; Kiyasseh et al., 2021; Gopal et al., 2021). A prominent example is SimCLR, which learns embeddings that pull together augmented views of the same instance while uniformly pushing apart other instances in the batch using the NT-Xent loss (Chen et al., 2020). Although originally developed for vision, CL has since been broadly adopted for representation learning on clinical time series (Weatherhead et al., 2022; Li and Gao, 2022; Zhang et al., 2022; Raghu et al., 2023; Noroozizadeh et al., 2023; Liu et al., 2023; King et al., 2024). However, existing methods treat all patients as equally strong negatives, ignoring that patients in clinical settings often share related diagnoses and disease profiles. While recent work has explored flat ICD code overlap as a similarity signal (Ding et al., 2024), this ignores the hierarchical structure of diagnosis taxonomies. We therefore leverage the full ICD ontology to derive fine-grained patient similarity and down-weight clinically similar negatives during contrastive pretraining, grounding the objective in clinically meaningful disease relationships.

2.2. Multimodal Learning for Clinical Time Series

Recent work shows that multimodal models can outperform unimodal baselines by leveraging complementary clinical signals (Khader et al., 2023; Jorf and Shamout, 2025). Accordingly, many approaches incorporate fusion mechanisms and report improved performance when combining physiologic time series with additional modalities, such as clinical notes (Lyu et al., 2023; Battula et al., 2024; Hooman et al., 2025). While effective, fusion-based approaches require all modalities to be available as inputs at deployment, an assumption that often fails in practice (Lang et al., 2025a,b; Fu et al., 2026). We instead study *training-time* multimodal supervision, leveraging auxiliary modality (e.g., clinical notes) during training while using only real-time vital signs at deployment for rapid prediction.

Within this paradigm, Ketabi and Ramachandram (2025) adopt CLIP-style contrastive pretraining to align physiologic signals with clinical notes, and subsequently fine-tune the resulting time-series encoder for downstream tasks. Ding et al. (2024) additionally regularize the learned representations by using a cross-entropy loss to match a diagnosis-based similarity distribution, alongside the CLIP objective. Alternative methods draw on privileged-information learning (Vapnik and Vashist, 2009; Chen et al., 2021), particularly modality synthesis, where auxiliary modalities available during training are generated at inference time to support prediction (Chartsias et al., 2017; Yu et al., 2018).

3. Methods

We define a two-stage framework consisting of ontology-aware contrastive pretraining followed by knowledge distillation fine-tuning. In the first stage (Section 3.1), we learn representations from vital-sign time series using CL that incorporates similarity scores derived from the ICD ontology. In the second stage (Section 3.2), we fine-tune the model using a knowledge distillation framework, where representations learned from clinical notes are distilled into the time-series encoder for downstream ICU prediction tasks.

3.1. Stage 1: Ontology-aware Contrastive Pretraining

Standard CL frameworks treat all negative pairs equally during training (Chen et al., 2020). However, in clinical settings, patients often share related disease profiles and should not be pushed arbitrarily far apart in the representation space. To address this limitation, we propose an ontology-aware CL approach that uses ICD-derived patient similarity to down-weight clinically similar negatives during pretraining. We begin by defining a measure of patient similarity based on the hierarchical structure of ICD ontology.

3.1.1. Patient Similarity Score Derivation

We obtain the ICD ontology from BioPortal (Noy et al., 2009) and use it to construct a hierarchical representation of diagnosis concepts. The resulting structure is a rooted tree: diagnoses share a common root and branch into progressively more specific categories (e.g., digestive diseases, circulatory diseases, congenital anomalies), with leaf nodes corresponding to individual ICD codes. The tree has maximum depth six and includes 14,870 nodes.

For two ICD codes a and b, let $\mathrm{Path}(a)$ denote the set of nodes on the path from the root to a, $\mathrm{Path}(b)$ denote the set of nodes on the path from the root to b, and define $\mathrm{Shared}(a,b) = |\mathrm{Path}(a) \cap \mathrm{Path}(b)| - 1$ as the number of nodes the two paths have in common excluding the root. We then define the similarity between a and b as the Jaccard overlap of their root-excluded path sets:

$$s(a,b) \;=\; \frac{\mathrm{Shared}(a,b)}{|\mathrm{Path}(a) \cup \mathrm{Path}(b)| - 1},$$

where $|\cdot|$ denotes set cardinality. This yields a similarity score that ranges from 0 (only the root is shared) to 1 (identical codes).

Patients typically have multiple diagnoses. For two patients A and B with diagnosis sets $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$, we compute patient-level similarity via a symmetric *best-match* aggregation over codes. Specifically, for each code in A, we take its maximum ontology similarity to any code in B and average across codes, and we repeat the same procedure in the reverse direction for B:

$$\mathrm{best}_A(a_i) = \max_{b \in B} s(a_i, b), \quad \mathrm{avg}_A = \frac{1}{m} \sum_{i=1}^{m} \mathrm{best}_A(a_i).$$

$$\mathrm{best}_B(b_j) = \max_{a \in A} s(a, b_j), \quad \mathrm{avg}_B = \frac{1}{n} \sum_{j=1}^{n} \mathrm{best}_B(b_j).$$

We then define the overall patient similarity as the mean of the two averages:

$$\mathrm{Sim}(A, B) = \frac{1}{2}\left(\mathrm{avg}_A + \mathrm{avg}_B\right).$$

3.1.2. Ontology-aware Contrastive Loss

We next use the derived similarity score to guide CL on vital-sign time series. Let $\mathbf{e}_i$ denote the ℓ_2-normalized representation of patient i produced by a time-series encoder f_{vitals}. Following standard CL practice, for each patient i in a batch, we construct a positive example $\mathbf{e}_i^+$ for each anchor $\mathbf{e}_i$ by applying transformations to the input. The representations of all other patients in the same batch are treated as negatives. The standard NT-Xent loss pushes all negative samples away from the anchor with equal force:

$$\mathcal{L}_{\text{NT-Xent}} = -\log \frac{\exp(\mathbf{e}_i^\top \mathbf{e}_i^+/\tau)}{\exp(\mathbf{e}_i^\top \mathbf{e}_i^+/\tau) + \sum_{j\neq i} \exp(\mathbf{e}_i^\top \mathbf{e}_j/\tau)},$$

where τ is a temperature parameter.

We then incorporate the patient similarity score from Section 3.1.1 into the contrastive loss. Intuitively, patients with highly overlapping diagnosis profiles should not be treated as equally strong negatives, since forcing their representations far apart can distort the embedding space. Concretely, we propose a weighted contrastive objective in which each negative pair (i, j) is assigned a weight $w_{ij} = \Phi(\mathrm{Sim}(A_i, A_j)) \in [0, 1]$.

$$\mathcal{L}^{(i)}_{\text{OW-NTXent}} = -\log \frac{\exp(\mathbf{e}_i^\top \mathbf{e}_i^+/\tau)}{\exp(\mathbf{e}_i^\top \mathbf{e}_i^+/\tau) + \sum_{j\neq i} w_{ij}\, \exp(\mathbf{e}_i^\top \mathbf{e}_j/\tau)}.$$

The transformation $\Phi : [0, 1] \to [0, 1]$ is a monotone decreasing transformation that controls the sharpness of down-weighting and can be instantiated by different parametric forms, including a power transform $\Phi(s) = (1-s)^\gamma$, an exponential decay $\Phi(s) = \exp(-\gamma s)$, or a hard threshold $\Phi(s) = \mathbb{I}[s < \delta]$. Larger γ or a smaller threshold δ yields a sharper weighting, more aggressively down-weighting highly similar pairs while placing relatively greater weight on dissimilar negatives.

3.2. Stage 2: Knowledge Distillation Finetuning

We next finetune the pretrained time series encoder using knowledge distillation to transfer complementary information from clinical notes. We first train a teacher model that uses clinical notes together with vital signs to generate informative soft targets (Section 3.2.1). We then finetune a student model initialized from the stage 1 pretrained encoder by distilling the teacher outputs while also optimizing the standard supervised loss (Section 3.2.2).

3.2.1. Teacher Model Training

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the training set, where $\mathbf{x}_i = [\mathbf{x}_i^{(orig)}, \mathbf{x}_i^{(vitals)}]$. Here, $\mathbf{x}_i^{(orig)}$ is the raw clinical note sequence and $\mathbf{x}_i^{(vitals)} \in \mathbb{R}^{T\times c}$ denotes the time-series vital signs with c channels over T time steps. The label y_i is the outcome of interest.

Clinical notes are often long and sparse, mixing salient information with noise. LLMs have demonstrated the ability to reason over complex information and generate concise

summaries for downstream tasks (Leong et al., 2024a; Zhu et al., 2025; Li et al., 2025; Lan et al., 2026a; Guo et al., 2026). Their capabilities have also been explored across a broad range of technical and biomedical applications (Huang et al., 2024; Leong et al., 2024b; Yuan et al., 2026; Nie et al., 2026; Li, 2026; Hu, 2026; Zhang et al., 2026a). Motivated by Battula et al. (2024), we augment the raw notes with LLM-generated summaries. Specifically, we use GPT-4o in a zero-shot setting to generate a concise summary $\mathbf{x}_i^{(summ)}$ for each raw note $\mathbf{x}_i^{(orig)}$. During training, we randomly select between the raw note and its summary with probability p and $1-p$, respectively:

$$\mathbf{x}_i^{(notes)} = \begin{cases} \mathbf{x}_i^{(orig)} & \text{with probability } p, \\ \mathbf{x}_i^{(summ)} & \text{with probability } 1-p, \end{cases}$$

For each example i, we encode each modality into a shared d-dimensional embedding space ($d = 768$) using modality-specific encoders:

$$\mathbf{h}_i^{(\text{notes})} = f_{\text{notes}}\left(\mathbf{x}_i^{(\text{notes})}\right) \in \mathbb{R}^d, \qquad \mathbf{h}_i^{(\text{vitals})} = f_{\text{vitals}}\left(\mathbf{x}_i^{(\text{vitals})}\right) \in \mathbb{R}^d.$$

We then fuse the modalities by element-wise addition:

$$\mathbf{h}_i^{(\text{fuse})} = \mathbf{h}_i^{(\text{notes})} + \mathbf{h}_i^{(\text{vitals})}.$$

The fused representation is passed to a classification head to produce teacher logits $\mathbf{z}_i^{(\text{teacher})} = g_{\text{teacher}}\left(\mathbf{h}_i^{(\text{fuse})}\right)$, followed by a softmax (multiclass) or sigmoid (binary) to obtain predicted probabilities $\hat{\mathbf{y}}_i^{(\text{teacher})}$.

We jointly train the encoders end-to-end using a hard supervised loss $\mathcal{L}_{\text{hard}}^{(i)}$ on ground truth labels. We use cross-entropy (CE) loss for multi-class outcomes and binary cross-entropy (BCE) for binary outcomes.

$$\mathcal{L}_{\text{hard}}^{(i)} = \begin{cases} \text{CE}\left(\hat{\mathbf{y}}_i^{(\text{teacher})}, y_i\right), & \text{multi-class}, \\ \text{BCE}\left(\hat{\mathbf{y}}_i^{(\text{teacher})}, y_i\right), & \text{binary}, \end{cases} \qquad \mathcal{L}_{\text{teacher}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\text{hard}}^{(i)}.$$

3.2.2. Student Model Training

After training the teacher model, we freeze its parameters and distill its learned knowledge into a student model that uses only time-series vital signs as input. For each example i, the student encoder produces a latent representation

$$\mathbf{h}_i^{(\text{student})} = f_{\text{vitals}}\left(\mathbf{x}_i^{(\text{vitals})}\right),$$

which is then passed to a classification head to produce student logits $\mathbf{z}_i^{(\text{student})} = g_{\text{student}}\left(\mathbf{h}_i^{(\text{student})}\right)$.

The student is trained with the same hard supervised loss $\mathcal{L}_{\text{hard}}^{(i)}$ and an additional soft distillation loss. For multi-class outcomes, we match predictions via KL-divergence:

$$\mathbf{p}_{i,T}^{(\text{teacher})} = \text{softmax}\left(\mathbf{z}_i^{(\text{teacher})}/T\right), \qquad \mathbf{p}_{i,T}^{(\text{student})} = \text{softmax}\left(\mathbf{z}_i^{(\text{student})}/T\right),$$
$$\mathcal{L}_{\text{distill}}^{(i)} = \text{KL}\left(\mathbf{p}_{i,T}^{(\text{teacher})} \,\|\, \mathbf{p}_{i,T}^{(\text{student})}\right).$$

For binary outcomes, we use the teacher's sigmoid output as a soft target:

$$p_{i,T}^{(\text{teacher})} = \sigma\left(z_i^{(\text{teacher})}/T\right), \quad p_{i,T}^{(\text{student})} = \sigma\left(z_i^{(\text{student})}/T\right),$$
$$\mathcal{L}_{\text{distill}}^{(i)} = \text{BCE}\left(p_{i,T}^{(\text{student})}, p_{i,T}^{(\text{teacher})}\right),$$

where T is a temperature parameter that softens the probability distributions, allowing the student to learn from the teacher's uncertainty.

The final objective is

$$\mathcal{L}_{\text{student}} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}_{\text{hard}}^{(i)} + \lambda_{\text{distill}}\,\mathcal{L}_{\text{distill}}^{(i)}\right),$$

where λ_{distill} is a hyperparameter controlling distillation strength. After training, the student model is deployed using only vital signs as input, enabling real-time predictions without requiring clinical notes.

4. Cohort and Experiment Setup

4.1. Cohort Selection and Benchmark Tasks

For our experiments, we construct time-series features for each ICU stay from the MIMIC-III benchmark dataset (Harutyunyan et al., 2019) and align them with the corresponding clinical notes (Johnson et al., 2016). We benchmark model performance on the following tasks widely used to evaluate ICU risk prediction models (Purushotham et al., 2018; Harutyunyan et al., 2019; Wang et al., 2020).

1. **In-hospital mortality prediction**: A binary classification task that predicts in-hospital mortality after the first T hours spent in the ICU. This task assesses the model's ability to identify high-risk patients early. We evaluate this task using the AUROC and the AUPRC.

2. **Length of stay prediction**: A 10-class classification task that predicts the remaining ICU length of stay from the first T hours of an ICU admission. Accurate length of stay prediction is important for hospital resource planning. We evaluate this task using the macro-averaged one-vs-rest AUROC and AUPRC.

For both tasks, we report results for $T = 48$ hours in Section 5 and additionally evaluate $T \in \{72, 96\}$ hours in Appendix C, where we observe similar performance trends.

4.2. Data Extraction and Feature Choices

We follow the pre-processing pipeline of prior work (Harutyunyan et al., 2019) and use 12 vital signs and physiological variables (diastolic blood pressure, fraction of inspired oxygen, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH). For all tasks, we discretize the time series into one-hour intervals and concatenate a binary missingness indicator for each variable, yielding a time-series representation $\mathbf{x}_i^{(\text{vitals})} \in \mathbb{R}^{T\times 24}$.

Observation horizon T	Training	Test
48 hours	16,861	3,055
72 hours	11,003	1,958
96 hours	7,788	1,418

Table 1: Dataset sizes (ICU stays) in the constructed cohort for each observation horizon.

We then align clinical notes to each ICU stay and retain only those recorded within the first T hours of the stay. We focus on nursing and radiology notes, the two most frequently available note types. For each stay, we sort notes by chart time and concatenate them into a single document. We then preprocess the text by lowercasing, removing bracketed de-identification spans, collapsing repeated separator characters, and normalizing whitespace. We additionally use GPT-4o to generate concise summaries of the clinical notes for data augmentation during teacher training. We accessed GPT-4o via the Microsoft Azure API, which is HIPAA-compliant and recommended by PhysioNet for use with MIMIC data. Appendix A reports ablation results comparing performance with and without LLM-generated summaries, showing consistent gains when summaries are combined with raw notes. Appendix B further shows that the multimodal teacher produces better-calibrated soft probabilities than a note-free teacher, motivating the use of distillation.

We split the dataset using the same train/test split as in Harutyunyan et al. (2019). We report final performance on the test set and compute 95% confidence intervals using 1,000 bootstrap resamples (Efron and Tibshirani, 1994). Table 1 summarizes the dataset sizes for each observation horizon T.

4.3. Dataset Choice and External Validation

Our experiments are conducted on MIMIC-III, which uniquely provides longitudinal clinical notes, specifically nursing notes written throughout a patient's ICU stay. Other commonly used benchmarks such as MIMIC-IV (Johnson et al., 2023) include only discharge summaries and radiology reports, making them less suitable for our setting where notes serve as a rich source of training-time supervision. To assess generalizability beyond MIMIC-III, we provide additional evaluation of our method on MIMIC-IV vital signs in Appendix F and eICU vital signs (Pollard et al., 2018) in Appendix G, where OC-Distill continues to show strong performance, demonstrating robustness across datasets.

4.4. Model Training and Implementation Details

To ensure a fair comparison, we use consistent model architectures across all methods and baselines. We parameterize the vitals encoder f_{vitals} as a BERT-style transformer (Devlin et al., 2019) with two layers and four attention heads. For clinical notes, we parameterize f_{notes} using Bio_ClinicalBERT (Alsentzer et al., 2019), a transformer-based language model shown to capture the terminology and structure of clinical documentation effectively.

For contrastive pretraining, positive pairs are constructed by applying transformations to the same vitals sequence. Specifically, given an input time series, we independently generate two views using three augmentations: (1) *Gaussian jitter*, which adds zero-mean

Gaussian noise to observed vital values; (2) *random time masking*, which masks random timesteps by setting all features at those timesteps to zero; and (3) *random feature masking*, which masks random vital-sign channels across all timesteps by setting their values to zero. We train with learning rate $\eta = 10^{-4}$, temperature $\tau = 1.0$, and batch size of 4096 for 1000 epochs. We use the power transformation $\Phi(s) = (1-s)^{\gamma}$ with $\gamma = 5$ as the similarity-weight transformation, and we study alternative choices of γ values and Φ in Section 5.1.3.

For knowledge distillation fine-tuning, we perform hyperparameter tuning via grid search over learning rate $\eta \in \{1 \times 10^{-4}, 5 \times 10^{-4}, 5 \times 10^{-5}\}$, distillation temperature $T \in \{1, 2, 5\}$, distillation loss weight $\lambda_{\text{distill}} \in \{1, 5, 10\}$, and the probability of using an LLM-generated note summary during training $p \in \{0, 0.5, 1.0\}$. Each model is trained for 100 epochs, and for each configuration we monitor validation AUROC throughout training and select the checkpoint with the best validation performance. The deployed student model contains approximately 19.3M parameters. On a single NVIDIA L40S GPU, contrastive pretraining takes 1.8 hours, the distillation pipeline takes 2.4 hours, and inference takes 0.0004 seconds per sample. All experiments were run on a single NVIDIA L40S GPU.

5. Results

In this section, we report results of OC-Distill and compare against a range of baselines. In Section 5.1, we evaluate ontology-aware contrastive pretraining, analyzing how ontology-derived similarities shape the embedding space and improve label efficiency. In Section 5.2, we evaluate the complete OC-Distill pipeline with distillation finetuning and show additional gains from incorporating clinical notes.

5.1. Contrastive Pretraining Results

Baselines. To quantify the benefit of leveraging the ICD ontology during contrastive pretraining, we compare our method against two baselines. (a) *SimCLR* (Chen et al., 2020) uses the standard NT-Xent objective and treats all non-matching instances as equally strong negatives (i.e., uniform weights $w_{ij} = 1$). (b) *Flat Diagnosis Match* (Ding et al., 2024) defines patient similarity via exact ICD-code overlap, computed as the number of shared ICD codes normalized by the total number of codes across the two patients. We then plug this flat similarity into the same reweighting formulation as our negative-weighted loss function, so any performance differences isolate the contribution of exploiting the ICD ontology.

5.1.1. Similarity Analysis and Representation Geometry

Similarity Analysis. We begin by examining how different similarity definitions translate into pairwise contrastive weights. As shown in Figure 2, flat diagnosis matching is inherently coarse. With thousands of ICD codes, exact matching overlooks the hierarchy and assigns zero similarity (weight = 1) to nearly half of patient pairs, effectively treating a large portion of negatives as equally dissimilar. In contrast, ontology-aware similarity yields a more informative measure that distributes weight across a continuum, with fewer than 4% of pairs assigned zero similarity. A power transformation of the ontology-aware similarity further sharpens this spectrum, shifting more weights toward 0 and enabling more aggressive reweighting during pretraining.

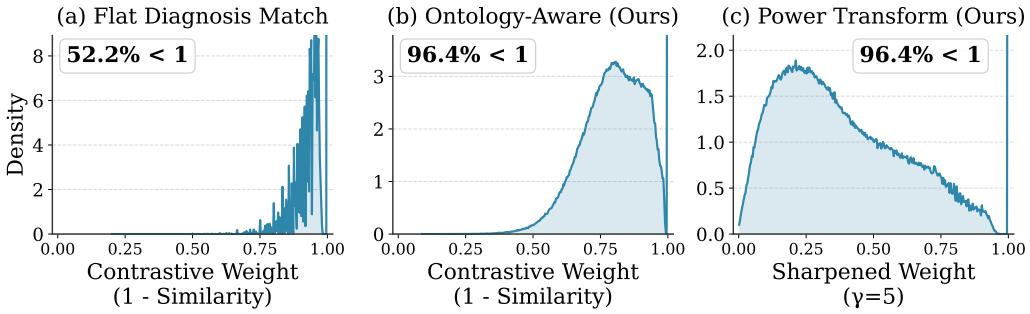

Figure 2: Distribution of contrastive weights for patient pairs. Flat diagnosis matching yields zero similarity (weight = 1) for half of pairs. Ontology-aware similarity produces a smoother distribution with gradations, while a power transformation sharpens the weights. Percentages indicate the fraction with weight < 1.

Embedding-space similarity analysis. We next evaluate whether the pretrained encoder produces embeddings that reflect the ontology-derived patient similarity structure. By incorporating ontology-aware weighting into the contrastive loss objective, patients that are close in embedding space should also be clinically similar in terms of diagnosis structure. We test this by comparing diagnosis similarity between nearest-neighbor pairs (K=5) in the learned embedding space and randomly sampled pairs. As shown in Figure 3, nearest-neighbor pairs exhibit significantly higher diagnosis similarity than random pairs (Mann–Whitney U: p <0.0001; r=0.198), validating that the representations capture diagnosis-relevant information. Additional results for different neighborhood sizes K are reported in Appendix D, confirming that the conclusion is robust to the choice of K.

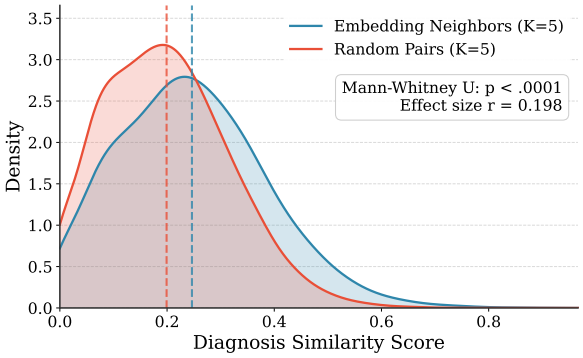

Figure 3: Diagnosis similarity distributions for embedding neighbors vs. random pairs. Embedding neighbors show higher similarity.

5.1.2. Pretraining Performance

Linear evaluation. We first assess the effect of contrastive pretraining via linear probing, where we freeze the encoder and train a linear classifier on top of the learned embeddings. Linear probing is most informative under label scarcity, and following standard protocol, we report results using 1%, 5%, and 10% labeled training data (Table 2). Compared to SimCLR, Flat Diagnosis CL consistently yields stronger performance, highlighting the value of incorporating diagnosis supervision beyond augmentations alone. Ontology-Aware CL further improves upon Flat Diagnosis CL and achieves the best results across tasks and all

Methods	Labels	In-Hospital Mortality		Length of Stay	
		AUROC	AUPRC	AUROC	AUPRC
SimCLR	1%	0.635 (0.603–0.666)	0.201 (0.172–0.238)	0.533 (0.519–0.547)	0.122 (0.119–0.130)
	5%	0.723 (0.695–0.750)	0.279 (0.235–0.328)	0.588 (0.572–0.602)	0.142 (0.138–0.151)
	10%	0.728 (0.701–0.754)	0.289 (0.244–0.340)	0.610 (0.595–0.624)	0.149 (0.144–0.158)
Flat Diagnosis CL	1%	0.647 (0.616–0.677)	0.208 (0.177–0.247)	0.536 (0.521–0.550)	**0.123 (0.120–0.131)**
	5%	0.731 (0.703–0.756)	0.294 (0.250–0.346)	0.591 (0.576–0.605)	0.142 (0.139–0.151)
	10%	0.741 (0.714–0.765)	0.301 (0.257–0.352)	0.610 (0.596–0.624)	0.150 (0.144–0.159)
Ontology-Aware CL (Ours)	1%	**0.673 (0.643–0.702)**	**0.230 (0.196–0.274)**	**0.541 (0.525–0.555)**	**0.123 (0.120–0.131)**
	5%	**0.750 (0.723–0.775)**	**0.319 (0.271–0.372)**	**0.600 (0.585–0.614)**	**0.144 (0.140–0.154)**
	10%	**0.757 (0.731–0.781)**	**0.328 (0.278–0.382)**	**0.616 (0.602–0.629)**	**0.152 (0.147–0.161)**

Table 2: Linear evaluation of contrastive pretraining methods with 1%, 5%, and 10% labeled training data. Ontology-aware CL consistently achieves the best results. Best is **bold**, second-best is underlined.

Methods	Labels	In-Hospital Mortality		Length of Stay	
		AUROC	AUPRC	AUROC	AUPRC
SimCLR	50%	0.730 (0.699–0.757)	0.332 (0.281–0.386)	0.668 (0.656–0.680)	0.163 (0.159–0.173)
	100%	0.781 (0.755–0.805)	0.377 (0.323–0.431)	0.676 (0.664–0.688)	0.170 (0.166–0.180)
Flat Diagnosis CL	50%	0.763 (0.735–0.788)	**0.360 (0.312–0.413)**	0.669 (0.657–0.682)	0.164 (0.160–0.174)
	100%	0.786 (0.760–0.809)	**0.387 (0.333–0.439)**	**0.678 (0.666–0.690)**	0.170 (0.166–0.180)
Ontology-Aware CL (Ours)	50%	**0.778 (0.752–0.802)**	0.359 (0.307–0.412)	**0.673 (0.659–0.685)**	**0.165 (0.161–0.175)**
	100%	**0.789 (0.764–0.812)**	0.379 (0.326–0.434)	0.676 (0.663–0.688)	**0.173 (0.168–0.184)**

Table 3: Full fine-tuning evaluation of contrastive pretraining models. Ontology-aware CL remains competitive and delivers strong performance across tasks. Best is **bold**, second-best is underlined.

label fractions, with the largest gains in the low-label regime. These findings indicate that leveraging ICD ontology structure produces more label-efficient representations.

Cross-time generalization. We further evaluate whether pretraining on $T = 48$ hour time-series features yields representations that generalize to downstream tasks with longer observation windows. Figure 4 reports linear probing performance evaluated at downstream horizons of $T = 48$, 72, and 96 hours. As expected, performance decreases for all methods as the downstream horizon increases. However, Ontology-Aware CL consistently remains the top-performing approach across observation horizons, demonstrating more robust cross-time generalization.

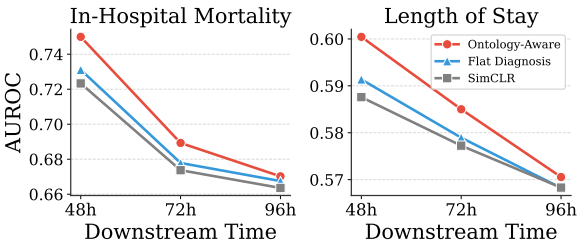

Figure 4: Cross-time generalization of linear-probe performance. Models were pretrained on 48-hour windows and evaluated at downstream horizons of 48, 72, and 96 hours.

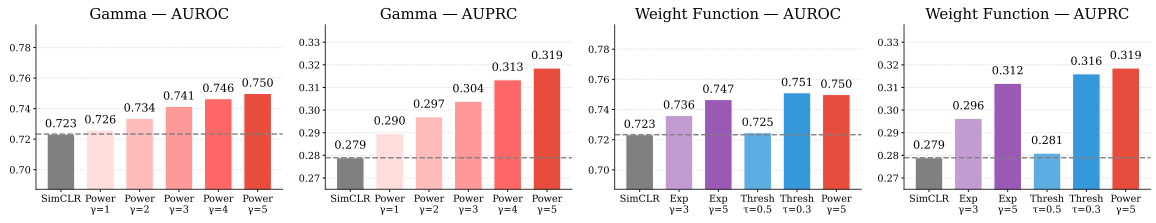

Figure 5: Linear probing performance under different weight transformations, evaluated using 5% labeled training data. Left pair: effect of the power-transform exponent γ. Right pair: comparison across weight functions. Similar trends are observed across other labeled-data fractions.

Full fine-tuning. We lastly evaluate pretraining performance under full fine-tuning. In this setting, we initialize the model from pretrained weights and update all parameters during supervised training. Following standard protocol, we evaluate with 50% and 100% labeled data, as full fine-tuning requires sufficient labeled examples to update all parameters reliably. Table 3 reports results, where Ontology-Aware CL remains strong across tasks.

5.1.3. Ablation Studies

Ablation effects of weight transformations. Our experiments incorporate similarity scores into the contrastive loss using the power transform $\Phi(s) = (1-s)^\gamma$ with $\gamma = 5$. We ablate this design by varying γ and the transformation family to quantify the effect of the transformation choice on performance. Figure 5 reports linear probing results under these settings. We observe that all transformations consistently outperform SimCLR. Within the power family, performance improves as γ increases, suggesting that more aggressive reweighting improves discriminability in supervised tasks. Exponential and threshold transforms show similar patterns, with sharpened settings (e.g., $\gamma = 5$, $\tau = 0.3$) achieving the best performance. Together, these results highlight the benefit of emphasizing clinically meaningful similarity structure during pretraining.

We further compare our simple Jaccard path-overlap similarity against established ontology similarity measures in Appendix E. Results show that our metric is strongly correlated with existing ontology similarity measures (Rada et al., 1989; Resnik, 1995; Jiang and Conrath, 1997), with higher correlation to edge-based than information-content-based alternatives. It also achieves competitive downstream performance, suggesting that it captures clinically meaningful ontology structure while remaining simple to compute.

5.2. End-to-End Evaluation of OC-Distill

In this section, we evaluate the complete OC-Distill pipeline end-to-end. Section 5.2.1 compares OC-Distill against unimodal and multimodal baselines, demonstrating state-of-the-art performance among methods that use only vital signs at inference. Section 5.2.2 validates the two-stage design through ablations, showing that each stage contributes meaningfully to the overall performance.

Method	Labels	In-Hospital Mortality		Length of Stay	
		AUROC	AUPRC	AUROC	AUPRC
Unimodal					
Transformer	50%	0.763 (0.736–0.788)	0.328 (0.280–0.381)	0.673 (0.661–0.684)	0.165 (0.160–0.175)
	100%	0.788 (0.762–0.811)	0.372 (0.320–0.422)	0.676 (0.664–0.688)	0.168 (0.164–0.179)
Multimodal (training time supervision)					
Cross-Modal Contrastive	50%	0.763 (0.736–0.788)	0.345 (0.295–0.401)	0.666 (0.654–0.678)	0.164 (0.159–0.175)
	100%	0.773 (0.746–0.798)	0.374 (0.326–0.425)	0.673 (0.662–0.684)	0.168 (0.164–0.179)
CKLE	50%	0.767 (0.741–0.791)	0.344 (0.297–0.397)	0.672 (0.661–0.685)	0.163 (0.159–0.174)
	100%	0.781 (0.754–0.805)	0.361 (0.313–0.414)	0.678 (0.667–0.690)	0.167 (0.163–0.177)
Synthetic Notes Augmentation	50%	0.759 (0.731–0.786)	0.334 (0.287–0.386)	0.665 (0.653–0.677)	0.161 (0.157–0.170)
	100%	0.774 (0.749–0.798)	0.359 (0.310–0.412)	0.675 (0.663–0.687)	0.165 (0.161–0.176)
OC-Distill (Ours)	50%	**0.776** (0.751–0.800)	**0.361** (0.311–0.414)	**0.676** (0.662–0.688)	**0.169** (0.164–0.180)
	100%	**0.793** (0.769–0.816)	**0.385** (0.333–0.439)	**0.682** (0.670–0.695)	**0.175** (0.171–0.186)

Table 4: Performance comparison between OC-Distill and baseline methods. OC-Distill consistently outperforms the unimodal baseline and other training time multimodal supervision baselines. Best is **bold**, second-best is underlined.

5.2.1. Performance Evaluations

Baselines. We compare our framework against a range of baselines proposed in the literature. We first consider **unimodal baselines** trained solely on vital signs. Specifically, we use a *Transformer-based architecture* (Vaswani et al., 2017), which leverages self-attention to model temporal dependencies and has been shown to outperform LSTMs on MIMIC benchmarks (Song et al., 2018). We also compare against **multimodal supervision baselines** that incorporate clinical notes during training only. These include: (a) *Cross-Modal Contrastive*, which aligns notes and vitals using a CLIP-style objective and then fine-tunes the vitals encoder (Ketabi and Ramachandram, 2025); (b) *CKLE*, which adds flat diagnosis similarity as an additional regularizer on top of the CLIP objective (Ding et al., 2024); and (c) *Synthetic Notes Augmentation*, which trains a note generator and augments the vitals predictor with synthesized notes, a strategy widely used in medical imaging (Chartsias et al., 2017; Yu et al., 2018).

Results. Table 4 summarizes the results. Overall, multimodal training outperforms unimodal baselines, demonstrating the benefit of incorporating clinical notes during training. Among multimodal baselines, synthetic note augmentation performs worse, suggesting that techniques effective in medical imaging may not transfer to clinical notes due to their noise and variability. Cross-modal contrastive training is competitive with CKLE, whereas OC-Distill consistently achieves the best performance across tasks and label fractions.

5.2.2. Ablation Studies

Effect of Stage 1 pretraining. We first compare OC-Distill with student models initialized from the Stage 1 pretrained encoder against those trained from scratch (Table 5).

Across both tasks and label fractions, pretrained initialization consistently improves performance. These results highlight the benefits of Ontology-Aware contrastive pretraining, suggesting that it learns representations that transfer effectively to downstream distillation and improve final performance.

		50% Labels		100% Labels	
Task	**Initialization**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
In-Hospital Mortality	Scratch	0.766	0.334	0.791	0.379
	Pretrained	**0.776**	**0.361**	**0.793**	**0.385**
Length of Stay	Scratch	0.672	0.162	0.680	0.172
	Pretrained	**0.676**	**0.169**	**0.682**	**0.175**

Table 5: Effect of Stage 1 pretraining on OC-Distill performance.

Effect of Stage 2 distillation. We further highlight the benefits of distillation from clinical notes by varying the distillation weight λ_{distill} (Figure 6). Across both tasks and metrics, enabling distillation ($\lambda_{\text{distill}} > 0$) consistently improves over the no-distillation baseline ($\lambda_{\text{distill}} = 0$). This highlights that teacher guidance from clinical notes provides complementary signal beyond supervised labels alone. Performance is strongest at moderate-to-large λ_{distill}, suggesting that distillation is most effective when soft targets contribute meaningfully to training without overwhelming the downstream objective.

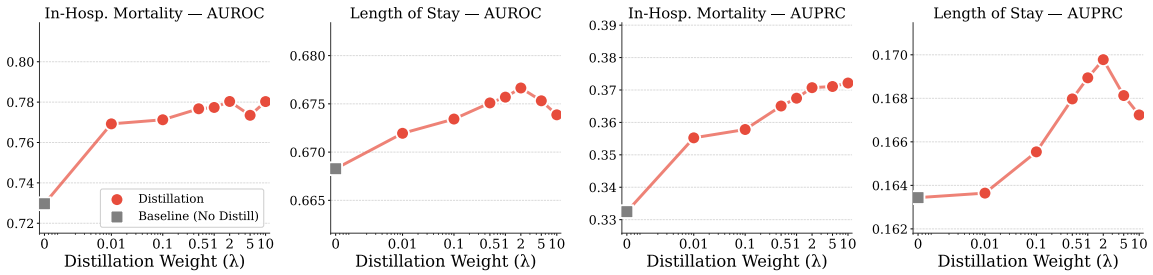

Figure 6: Effect of λ_{distill} on model performance, evaluated using 50% labeled training data. Across tasks, distillation consistently improves over the no-distillation baseline. Similar trends are observed across other labeled-data fractions.

6. Discussion

We propose OC-Distill, a two-stage framework that improves downstream ICU risk performance on physiological time-series data. By leveraging the hierarchical structure of the ICD ontology, our approach derives more informative patient similarity measures and incorporates them into contrastive pretraining to better shape the embedding space. Building on the pretrained representations, we further demonstrate that cross-modal knowledge distillation

from clinical notes provides complementary supervision, yielding additional performance gains while enabling deployment with real-time vital signs alone.

This work has several strengths. First, the framework demonstrates consistent gains across multiple tasks, label fractions, and observation horizons. We further conduct paired bootstrap testing to assess statistical significance. In the low-label setting (Table 2), OC-Distill significantly outperforms both baselines on mortality across all label fractions and metrics (all $p < 0.05$), and on length-of-stay in 5 of 6 AUROC comparisons. For the full pipeline evaluation (Table 4), significant gains persist at the 50% label fraction over *Cross-Modal Contrastive* and *Synthetic Notes Augmentation*. These results confirm significant gains, especially in the low-label regime where clinical data are often limited. Second, unlike large pretrained EHR models that often require rich inputs and substantial compute at deployment, the student model operates on vital signs alone at inference, making OC-Distill lightweight and practical for real-time ICU monitoring. Third, the ontology-aware reweighting scheme is loss-agnostic and can in principle be applied to other contrastive objectives beyond NT-Xent (Khosla et al., 2020), making it broadly applicable to other contrastive pretraining settings. Lastly, while we demonstrate the distillation stage with clinical notes, the teacher-student framework is modality-agnostic and naturally extends to other modalities or combinations of modalities.

This study also has its own limitations. First, our patient similarity measure depends on ICD coding, which may not fully capture clinical similarity due to noise, missing codes, or systematic biases such as upcoding (Silverman and Skinner, 2004; Coustasse, 2021). Future work could incorporate additional information such as laboratory measurements, medications, or procedures to define more robust notions of patient similarity. Second, our framework currently relies solely on clinical notes as the complementary modality. Other modalities such as medical imaging or structured EHR variables may capture complementary information not present in notes (Fu et al., 2021; Shen et al., 2021; Hayat et al., 2022; Wu et al., 2025; Zhang et al., 2025b), and exploring richer multimodal supervision remains an important direction for future work (Chen et al., 2025; Zhu et al., 2026). Third, our evaluation focuses on ICU vital signs, and it remains an open question whether the framework generalizes to other clinical time series and physiological waveform data (Tang et al., 2024; Oh and Bui, 2025; Zhang et al., 2025a, 2026b). Lastly, while OC-Distill improves predictive performance, our current analysis does not fully characterize which clinical concepts are transferred from clinical notes to vital-sign representations during distillation. Future work could further investigate which note-derived clinical concepts drive predictive performance through interpretability methods (Lundberg and Lee, 2017; Singh et al., 2024; Liang et al., 2025a).

Acknowledgments

This work was supported by funds from Samsung Electronics Co., Ltd.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint*

arXiv:1904.03323, 2019.

Harshavardhan Battula, Jiacheng Liu, and Jaideep Srivastava. Enhancing in-hospital mortality prediction using multi-representational learning with llm-generated expert summaries. *arXiv preprint arXiv:2411.16818*, 2024.

Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Multimodal mr synthesis via modality-invariant latent representation. *IEEE transactions on medical imaging*, 37(3):803–814, 2017.

Cheng Chen, Qi Dou, Yueming Jin, Quande Liu, and Pheng Ann Heng. Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE transactions on medical imaging*, 41(3):621–632, 2021.

Lei Chen, Kyoungsuk Park, and Junetae Kim. Multimodal contrastive learning with early fusion for robust medical signal representation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4649–4653, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Alberto Coustasse. Upcoding medicare: is healthcare fraud and abuse increasing? *Perspectives in health information management*, 18(4):1f, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Sirui Ding, Jiancheng Ye, Xia Hu, and Na Zou. Distilling the knowledge from large-language model for health event prediction. *Scientific Reports*, 14(1):30675, 2024.

Yongkang Ding, Yuxiang Wang, Yiyun Su, Yu Tian, Zi Ye, and Xiangzhou Jian. Synergistic cross-modal prompt learning and structural knowledge distillation for cloth-changing person re-identification. *Knowledge-Based Systems*, 345:116132, 2026. doi: 10.1016/j.knosys.2026.116132.

Hongbo Du, Zixin Lu, and Jiaming Qu. Possible or definite? a benchmark for evaluating diagnostic uncertainty preservation in clinical text. *arXiv preprint arXiv:2606.18471*, 2026.

Wenchang Duan. Maven-t: Multi-agent environment-aware enhanced neural trajectory predictor with reinforcement learning. *arXiv preprint arXiv:2604.10169*, 2026.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Rong Fu, Ziming Wang, Chunlei Meng, Jiaxuan Lu, Jiekai Wu, Kangan Qian, Hao Zhang, and Simon Fong. Missing-by-design: Certifiable modality deletion for revocable multimodal sentiment analysis. *arXiv preprint arXiv:2602.16144*, 2026.

Xinmei Fu, Zhenzhou Shao, Ying Qu, Yong Guan, Yibo Zou, Zhiping Shi, and Jindong Tan. Fast and unsupervised non-local feature learning for direct volume rendering of 3d medical images. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5886–5891. IEEE, 2021.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine learning for health*, pages 156–167. PMLR, 2021.

Xingang Guo, Yaxin Li, Xiangyi Kong, Yilan Jiang, Xiayu Zhao, Zhihua Gong, Yufan Zhang, Daixuan Li, Tianle Sang, Beixiao Zhu, et al. Toward engineering agi: Benchmarking the engineering design capabilities of llms. *Advances in Neural Information Processing Systems*, 38, 2026.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Nikkie Hooman, Zhongjie Wu, Eric C Larson, and Mehak Gupta. Equitable electronic health record prediction with fame: Fairness-aware multimodal embedding. *arXiv preprint arXiv:2506.13104*, 2025.

Yihan Hu. Toward retrieval-grounded evaluation for conversational large language model–based risk assessment. *JMIR AI*, 5:e90759, 2026.

Yihan Hu, Jingyuan Han, and Mingxin Liu. Protocol-aware epidemic forecasting across heterogeneous public health surveillance systems. *Frontiers in Public Health*, 14:1829302, 2026.

Yuanhao Huang, Zhaowei Han, Xin Luo, Xuteng Luo, Yijia Gao, Meiqi Zhao, Feitong Tang, Yiqun Wang, Jiyu Chen, Chengfan Li, et al. Building a literature knowledge base towards transparent biomedical ai. *bioRxiv*, pages 2024–09, 2024.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th research on computational linguistics international conference*, pages 19–33, 1997.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

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.

Baraa Al Jorf and Farah Shamout. Medpatch: Confidence-guided multi-stage fusion for multimodal clinical data. *arXiv preprint arXiv:2508.09182*, 2025.

Sara Ketabi and Dhanesh Ramachandram. Bridging electronic health records and clinical texts: Contrastive learning for enhanced clinical tasks. *arXiv preprint arXiv:2505.17643*, 2025.

Firas Khader, Jakob Nikolas Kather, Gustav Müller-Franzes, Tianci Wang, Tianyu Han, Soroosh Tayebi Arasteh, Karim Hamesch, Keno Bressem, Christoph Haarburger, Johannes Stegmaier, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports*, 13(1):10666, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Ryan King, Shivesh Kodali, Conrad Krueger, Tianbao Yang, and Bobak J Mortazavi. An efficient contrastive unimodal pretraining method for ehr time series data. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8. IEEE, 2024.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

Guangchen Lan, Lian Xiong, Xin Zhou, Hejie Cui, Yuwei Zhang, Mao Li, Zhenyu Shi, Besnik Fetahu, Lihong Li, and Xian Li. Alternating reinforcement learning with contextual rubric rewards: Beyond the scalarization strategy. *arXiv preprint arXiv:2603.15646*, 2026a.

Qizhen Lan and Qing Tian. Acam-kd: adaptive and cooperative attention masking for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3957–3966, 2025.

Qizhen Lan, Yu-Chun Hsu, Nida Saddaf Khan, and Xiaoqian Jiang. Reco-kd: Region-and context-aware knowledge distillation for efficient 3d medical image segmentation. *arXiv preprint arXiv:2601.08301*, 2026b.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18035–18043, 2025a.

Jian Lang, Rongpei Hong, Zhangtao Cheng, Ting Zhong, Yong Wang, and Fan Zhou. Redeeming modality information loss: Retrieval-guided conditional generation for severely modality missing learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1241–1252, 2025b.

Hui Yi Leong, Yifan Gao, and Shuai Ji. A gen ai framework for medical note generation. In *2024 6th international conference on artificial intelligence and computer applications (ICAICA)*, pages 423–429. IEEE, 2024a.

Huiyi Leong, Yifan Gao, Shuai Ji, Yang Zhang, and Uktu Pamuksuz. Efficient fine-tuning of large language models for automated medical documentation. In *2024 4th International Conference on Digital Society and Intelligent Systems (DSInS)*, pages 204–209. IEEE, 2024b.

Linjun Li. Same dangerous objective, opposite advice: Direct exposure versus multi-agent mediation. *arXiv preprint arXiv:2607.21518*, 2026.

Rui Li and Jing Gao. Multi-modal contrastive learning for healthcare data analytics. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 120–127. IEEE, 2022.

Songze Li, Zhiqiang Liu, Zhengke Gui, Huajun Chen, and Wen Zhang. Enrich-on-graph: Query-graph alignment for complex reasoning with llm enriching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7683–7703, 2025.

Zhongyuan Liang, Zachary T Rewolinski, Abhineet Agarwal, Tiffany M Tang, and Bin Yu. Local mdi+: Local feature importances for tree-based models. *arXiv preprint arXiv:2506.08928*, 2025a.

Zhongyuan Liang, Arvind Suresh, and Irene Y. Chen. Treatment non-adherence bias in clinical machine learning: A real-world study on hypertension medication. In *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pages 430–442. PMLR, 2025b.

Ziyu Liu, Azadeh Alavi, Minyi Li, and Xiang Zhang. Self-supervised contrastive learning for medical time series: A systematic review. *Sensors*, 23(9):4221, 2023.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2022, page 719, 2023.

Chunlei Meng, Guanhong Huang, Rong Fu, Runmin Jian, Zhongxue Gan, and Chun Ouyang. Clcr: cross-level semantic collaborative representation for multimodal learning. *arXiv preprint arXiv:2602.19605*, 2026.

Allen Nie, Xavier Daull, Zhiyi Kuang, Abhinav Akkiraju, Anish Chaudhuri, Max Piasevoli, Ryan Rong, YuCheng Yuan, Prerit Choudhary, Shannon Xiao, et al. Understanding the challenges in iterative generative optimization with llms. *arXiv preprint arXiv:2603.23994*, 2026.

Shahriar Noroozizadeh, Jeremy C Weiss, and George H Chen. Temporal supervised contrastive learning for modeling patient risk progression. In *Machine Learning for Health (ML4H)*, pages 403–427. PMLR, 2023.

Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl_2):W170–W173, 2009.

YongKyung Oh and Alex Bui. Multi-view contrastive learning for robust domain adaptation in medical time series analysis. *arXiv preprint arXiv:2506.22393*, 2025.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178, 2018.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, 19 (1):17–30, 1989.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Sequential multi-dimensional self-supervised learning for clinical time series. In *International Conference on Machine Learning*, pages 28531–28548. PMLR, 2023.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

Tom M Seinen, Jan A Kors, Erik M van Mulligen, and Peter R Rijnbeek. Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27:e66910, 2025.

Maohao Shen, Jacky Y Zhang, Leihao Chen, Weiman Yan, Neel Jani, Brad Sutton, and Oluwasanmi Koyejo. Labeling cost sensitive batch active learning for brain tumor segmentation. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1269–1273. IEEE, 2021.

Satya Narayan Shukla and Benjamin M Marlin. Integrating physiological time series and clinical notes with deep learning for improved icu mortality prediction. *arXiv preprint arXiv:2003.11059*, 2020.

Elaine Silverman and Jonathan Skinner. Medicare upcoding and hospital ownership. *Journal of health economics*, 23(2):369–389, 2004.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

Huan Song, Deepta Rajan, Jayaraman Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

Weitao Tang, Nhi Tran, Nasim Katebi, Reza Sameni, Gari D Clifford, David Walker, Vaishnavi Horlali, Callum Taylor, Robert Galinsky, and Faezeh Marzbanrad. Advancing fetal surveillance with physiological sensing: Detecting hypoxia in fetal sheep. In *2024 IEEE SENSORS*, pages 1–4. IEEE, 2024.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.

Yuqing Wang, Yun Zhao, Rachael Callcut, and Linda Petzold. Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. *arXiv preprint arXiv:2203.14469*, 2022.

Addison Weatherhead, Robert Greer, Michael-Alice Moga, Mjaye Mazwi, Danny Eytan, Anna Goldenberg, and Sana Tonekaboni. Learning unsupervised representations for icu timeseries. In *Conference on Health, Inference, and Learning*, pages 152–168. PMLR, 2022.

Hao Wu, Hui Li, and Yiyun Su. Bridging the perception-cognition gap: Re-engineering sam2 with hilbert-mamba for robust vlm-based medical diagnosis. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4275–4278. IEEE, 2025.

Biting Yu, Luping Zhou, Lei Wang, Jurgen Fripp, and Pierrick Bourgeat. 3d cgan based cross-modality mr image synthesis for brain tumor segmentation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 626–630. IEEE, 2018.

Yucheng Yuan, Yuanfeng Ji, Zhongxiao Li, and Ruijiang Li. Sp-mind: An autonomous reasoning agent for spatial proteomics analysis. *arXiv preprint arXiv:2606.24235*, 2026.

Yike Zhang, Zuodong Xiang, and Hailu Xu. Performance-efficiency trade-offs in human preference prediction: A comparative study of traditional machine learning and large language models. In *2026 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2026a.

Yue Zhang, Cynthia J Roberts, Swati Padhee, Nathan Doble, Srinivasan Parthasarathy, and Phillip Thomas Yuhas. Machine learning detection of keratoconus through ocular response analyzer waveform parameters. *Investigative Ophthalmology & Visual Science*, 66(8):3378–3378, 2025a.

Yue Zhang, Swati Padhee, Phillip T Yuhas, Cynthia J Roberts, and Srinivasan Parthasarathy. Adaptive temporal mixture of experts for predicting stiffness metrics from the ocular response analyzer and identifying keratoconus: Stiffness estimates from ocular response analyzer. *American journal of ophthalmology*, 2026b.

Zhenhao Zhang, Xinyu Bao, Linghua Jiang, Xin Luo, Zheyu Zhang, Jing Yin, Meiqi Zhao, Yichun Wang, Annelise Comai, Joerg Waldhaus, et al. Developing a general ai model for integrating diverse genomic modalities and comprehensive genomic knowledge. *Nucleic acids research*, 53(21):gkaf1269, 2025b.

Ziqi Zhang, Chao Yan, Xinmeng Zhang, Steve L Nyemba, and Bradley A Malin. Forecasting the future clinical events of a patient through contrastive learning. *Journal of the American Medical Informatics Association*, 29(9):1584–1592, 2022.

Wenjie Zhu, Yabin Zhang, Xin Jin, Wenjun Zeng, and Lei Zhang. Ants: Adaptive negative textual space shaping for ood detection via test-time mllm understanding and reasoning. *arXiv preprint arXiv:2509.03951*, 2025.

Wenjie Zhu, Yabin Zhang, Liang Xu, Xin Jin, Wenjun Zeng, and Lei Zhang. Dual distribution estimation for zero-shot noisy test-time adaptation with vlms. *arXiv preprint arXiv:2606.25758*, 2026.

Appendix A. Teacher Summary Fraction Ablation

Our experiments use GPT-4o summaries as additional context when training the teacher model. In this appendix, we ablate the fraction of summaries included in the teacher inputs, and report teacher model performance averaged over the 50% and 100% label fractions. Table 6 summarizes the results. We observe that a 50% summary fraction consistently yields the best performance, suggesting that incorporating LLM-generated summaries during training provides additional signal without overwhelming the original inputs.

Task	Fraction of LLM Summary	AUROC	AUPRC
In-Hospital Mortality	0%	0.872 (0.852–0.890)	0.512 (0.459–0.565)
	50%	**0.879** (0.859–0.897)	**0.527** (0.473–0.579)
	100%	0.856 (0.836–0.874)	0.438 (0.386–0.495)
Length of Stay	0%	0.696 (0.684–0.707)	0.180 (0.175–0.191)
	50%	**0.704** (0.693–0.716)	**0.185** (0.180–0.198)
	100%	0.670 (0.658–0.682)	0.162 (0.159–0.171)

Table 6: Teacher model performance with varying fractions of GPT-4o summaries, averaged across the 50% and 100% label fractions. Best is **bold**.

Appendix B. Teacher Calibration Analysis

In this section, we evaluate whether incorporating clinical notes improves the calibration of the teacher's soft probabilities, further motivating the use of soft-logit distillation. As shown in Table 7, the multimodal teacher achieves lower expected calibration error (ECE) and Brier score than the note-free teacher across both tasks. This suggests that the teacher's soft probabilities provide additional supervision signals enriched by clinical notes that can be distilled to the student.

Teacher	In-Hospital Mortality		Length of Stay	
	ECE	Brier	ECE	Brier
Note-free	0.062	0.171	0.045	0.751
With notes	**0.043**	**0.097**	**0.029**	**0.728**

Table 7: Teacher calibration with and without clinical notes.

Appendix C. Supplementary Results with Longer Observation Horizons

The main paper reports results using an observation horizon of $T = 48$ hours. In this appendix, we evaluate the robustness of the method when training with longer horizons, $T \in \{72, 96\}$ hours. Tables 8 and 9 present linear evaluation results for contrastive pretraining, while Tables 10 and 11 report full fine-tuning performance with knowledge distillation. The results at $T \in \{72, 96\}$ hours follow the same trends as the main paper. Our method performs strongly across tasks and stays competitive across all settings.

Methods	Labels	In-Hospital Mortality		Length of Stay	
		AUROC	AUPRC	AUROC	AUPRC
SimCLR	1%	0.641 (0.606–0.677)	0.228 (0.194–0.272)	0.576 (0.560–0.591)	0.130 (0.126–0.140)
	5%	0.675 (0.643–0.707)	0.265 (0.223–0.314)	0.588 (0.570–0.606)	0.145 (0.139–0.156)
	10%	0.736 (0.707–0.763)	**0.313 (0.267–0.369)**	0.602 (0.586–0.619)	0.149 (0.144–0.161)
Flat Diagnosis CL	1%	0.662 (0.626–0.700)	0.254 (0.215–0.305)	0.579 (0.563–0.594)	0.132 (0.128–0.141)
	5%	0.674 (0.642–0.706)	0.261 (0.220–0.312)	0.594 (0.577–0.612)	**0.145 (0.140–0.157)**
	10%	0.731 (0.700–0.760)	**0.313 (0.267–0.369)**	0.610 (0.593–0.627)	0.149 (0.143–0.160)
Ontology-Aware CL (Ours)	1%	**0.681 (0.649–0.715)**	**0.264 (0.225–0.314)**	**0.585 (0.569–0.600)**	**0.134 (0.130–0.145)**
	5%	**0.694 (0.663–0.725)**	**0.274 (0.231–0.326)**	**0.600 (0.582–0.617)**	0.144 (0.139–0.155)
	10%	**0.738 (0.709–0.765)**	**0.313 (0.267–0.369)**	**0.616 (0.599–0.631)**	**0.150 (0.145–0.162)**

Table 8: Linear evaluation of contrastive pretraining methods with a 72-hour observation horizon. Best is **bold**, second-best is underlined.

Methods	Labels	In-Hospital Mortality		Length of Stay	
		AUROC	AUPRC	AUROC	AUPRC
SimCLR	1%	0.619 (0.580–0.661)	0.238 (0.199–0.290)	0.563 (0.545–0.579)	0.128 (0.123–0.140)
	5%	0.676 (0.635–0.712)	0.282 (0.235–0.340)	0.578 (0.560–0.596)	0.143 (0.137–0.157)
	10%	0.704 (0.667–0.739)	**0.310 (0.258–0.368)**	0.575 (0.558–0.593)	0.144 (0.138–0.159)
Flat Diagnosis CL	1%	0.618 (0.579–0.660)	**0.239 (0.199–0.293)**	0.564 (0.548–0.581)	0.128 (0.124–0.141)
	5%	0.670 (0.629–0.708)	0.278 (0.233–0.337)	0.577 (0.559–0.596)	0.143 (0.137–0.157)
	10%	0.698 (0.661–0.734)	0.303 (0.254–0.364)	0.575 (0.557–0.592)	0.146 (0.139–0.161)
Ontology-Aware CL (Ours)	1%	**0.621 (0.582–0.662)**	0.237 (0.200–0.289)	**0.565 (0.548–0.582)**	**0.130 (0.126–0.144)**
	5%	**0.685 (0.646–0.721)**	**0.289 (0.243–0.343)**	**0.584 (0.566–0.602)**	**0.147 (0.141–0.162)**
	10%	**0.714 (0.678–0.746)**	0.309 (0.260–0.367)	**0.587 (0.570–0.606)**	**0.148 (0.142–0.163)**

Table 9: Linear evaluation of contrastive pretraining methods with a 96-hour observation horizon. Best is **bold**, second-best is underlined.

Method	Labels	In-Hospital Mortality		Length of Stay	
		AUROC	AUPRC	AUROC	AUPRC
Transformer	50%	0.719 (0.689–0.748)	0.285 (0.246–0.334)	0.649 (0.634–0.662)	0.162 (0.156–0.173)
	100%	0.725 (0.697–0.754)	0.293 (0.251–0.348)	0.659 (0.645–0.672)	0.169 (0.163–0.181)
Cross-Modal Contrastive	50%	0.725 (0.697–0.757)	0.312 (0.267–0.370)	0.641 (0.627–0.656)	0.162 (0.157–0.175)
	100%	0.737 (0.710–0.767)	0.320 (0.275–0.382)	0.649 (0.634–0.663)	0.165 (0.159–0.177)
CKLE	50%	0.714 (0.682–0.744)	0.279 (0.240–0.326)	0.652 (0.638–0.665)	0.161 (0.157–0.172)
	100%	0.733 (0.704–0.762)	0.320 (0.275–0.374)	**0.661** (0.648–0.674)	**0.170** (0.165–0.183)
Synthetic Notes Augmentation	50%	0.720 (0.691–0.748)	0.281 (0.243–0.332)	0.639 (0.625–0.653)	0.157 (0.152–0.168)
	100%	0.737 (0.706–0.767)	0.326 (0.279–0.384)	0.654 (0.640–0.668)	0.167 (0.162–0.180)
OC-Distill (Ours)	50%	**0.751** (0.723–0.780)	**0.350** (0.304–0.411)	**0.654** (0.641–0.667)	**0.164** (0.159–0.175)
	100%	**0.770** (0.742–0.798)	**0.372** (0.323–0.433)	0.657 (0.643–0.671)	0.166 (0.162–0.178)

Table 10: Performance comparison between OC-Distill and baseline methods with a 72-hour observation horizon. Best is **bold**, second-best is underlined.

Method	Labels	In-Hospital Mortality		Length of Stay	
		AUROC	AUPRC	AUROC	AUPRC
Transformer	50%	0.651 (0.613–0.686)	0.251 (0.213–0.305)	0.627 (0.609–0.644)	0.160 (0.153–0.173)
	100%	0.693 (0.655–0.728)	0.309 (0.256–0.368)	0.638 (0.621–0.654)	0.164 (0.157–0.179)
Cross-Modal Contrastive	50%	0.533 (0.491–0.572)	0.175 (0.148–0.209)	0.561 (0.542–0.579)	0.129 (0.125–0.141)
	100%	0.573 (0.532–0.611)	0.186 (0.160–0.225)	0.630 (0.611–0.647)	0.159 (0.151–0.176)
CKLE	50%	0.663 (0.625–0.699)	0.279 (0.234–0.341)	0.629 (0.612–0.645)	0.155 (0.150–0.167)
	100%	0.701 (0.665–0.737)	0.307 (0.258–0.363)	0.639 (0.621–0.655)	0.165 (0.158–0.178)
Synthetic Notes Augmentation	50%	0.675 (0.637–0.711)	0.271 (0.229–0.327)	0.628 (0.611–0.644)	0.158 (0.152–0.170)
	100%	0.736 (0.699–0.769)	0.335 (0.283–0.398)	0.631 (0.614–0.648)	0.161 (0.155–0.178)
OC-Distill (Ours)	50%	**0.731** (0.694–0.764)	**0.358** (0.302–0.419)	**0.641** (0.623–0.656)	**0.165** (0.159–0.180)
	100%	**0.742** (0.709–0.775)	**0.350** (0.295–0.413)	**0.639** (0.621–0.654)	**0.165** (0.159–0.178)

Table 11: Performance comparison between OC-Distill and baseline methods with a 96-hour observation horizon. Best is **bold**, second-best is underlined

Appendix D. Sensitivity to the choice of K in neighbor analysis

We repeated the embedding-space neighbor analysis with $K = 1$ and $K = 3$ in addition to the main-text setting of $K = 5$. As shown in Table 12, the conclusion is stable across all choices of K: nearest-neighbor pairs consistently exhibit higher diagnosis similarity than randomly sampled pairs, and all differences remain statistically significant.

K	KNN Mean	Random Mean	Effect Size r	Mann–Whitney p
1	0.251	0.200	0.214	< 0.0001
3	0.248	0.200	0.205	< 0.0001
5	0.246	0.199	0.198	< 0.0001

Table 12: Sensitivity of the embedding-space similarity analysis to the choice of neighborhood size K.

Appendix E. Comparison with Established Ontology Similarity Metrics

We compare our Jaccard path-overlap similarity against established ontology similarity measures commonly used in the biomedical ontology literature, including edge-based and information-content-based metrics (Rada et al., 1989; Resnik, 1995; Jiang and Conrath, 1997). We first assess whether our simple ICD path-overlap metric captures a structure similar to that of established ontology measures and then evaluate whether the choice of ontology similarity metric affects downstream performance.

Table 13 reports Spearman correlations between our similarity metric and these established alternatives. Results show that our metric is highly correlated with all three measures, with the strongest correlation to the edge-based metric as expected. Table 14 reports linear evaluation performance. Our simple metric achieves competitive or best performance

across label fractions, suggesting that a simple Jaccard metric captures the clinically relevant structure needed for contrastive pretraining while avoiding additional complexity.

Metric	**Spearman ρ**
Rada (Rada et al., 1989)	0.950
Jiang–Conrath (Jiang and Conrath, 1997)	0.934
Resnik (Resnik, 1995)	0.840

Table 13: Correlation between our ICD path-overlap similarity and established ontology similarity metrics.

Metric	**AUROC**			**AUPRC**		
	1%	**5%**	**10%**	**1%**	**5%**	**10%**
Resnik (Resnik, 1995)	0.651	0.735	0.744	0.210	0.298	0.305
Jiang–Conrath (Jiang and Conrath, 1997)	0.669	0.745	0.755	0.225	0.311	0.319
Rada (Rada et al., 1989)	0.669	**0.751**	**0.758**	0.229	**0.321**	**0.332**
Ours	**0.673**	0.750	**0.758**	**0.230**	0.319	0.328

Table 14: Downstream performance using different ontology similarity metrics in contrastive pretraining. Results are reported for linear evaluation under 1%, 5%, and 10% labeled training data. Best results are **bold**.

Appendix F. Validation on MIMIC-IV

To assess the generalizability of OC-Distill beyond MIMIC-III, we evaluate the student model on vital signs from MIMIC-IV (Johnson et al., 2023). Since our framework requires only vital signs at inference, the student model can be directly applied to MIMIC-IV without any retraining. We follow the same preprocessing pipeline and benchmark tasks as in Section 4. Table 15 reports results, where OC-Distill achieves strong performance across both tasks.

	In-Hospital Mortality		**Length of Stay**	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Transformer	0.792 (0.775–0.809)	0.368 (0.333–0.408)	0.668 (0.658–0.676)	0.158 (0.155–0.165)
Cross-Modal Contrastive	0.792 (0.775–0.809)	0.361 (0.326–0.399)	0.670 (0.661–0.680)	**0.162 (0.158–0.170)**
CKLE	0.797 (0.780–0.813)	0.357 (0.322–0.394)	0.668 (0.658–0.677)	0.157 (0.153–0.163)
Synthetic Notes Augmentation	0.794 (0.777–0.811)	0.367 (0.331–0.406)	0.665 (0.655–0.674)	0.154 (0.151–0.161)
OC-Distill (Ours)	**0.813 (0.797–0.828)**	**0.380 (0.346–0.419)**	**0.672 (0.663–0.682)**	0.158 (0.155–0.165)

Table 15: Validation on MIMIC-IV vital signs. OC-Distill demonstrates strong performance across both tasks. Best results are **bold**.

Appendix G. Validation on eICU

To further assess the generalizability of OC-Distill beyond MIMIC-III, we evaluate the student model on vital signs from the eICU dataset (Pollard et al., 2018). Tables 16 and 17 report results for in-hospital mortality and length-of-stay prediction. OC-Distill achieves strong performance for mortality prediction and remains competitive for length-of-stay prediction across horizons.

Method	48h		72h		96h	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Transformer	0.714 (0.708–0.720)	0.262 (0.252–0.271)	0.702 (0.694–0.709)	0.273 (0.263–0.284)	0.660 (0.650–0.669)	0.255 (0.245–0.267)
Cross-Modal Contrastive	0.707 (0.701–0.714)	0.255 (0.245–0.264)	0.696 (0.689–0.704)	0.256 (0.246–0.267)	0.591 (0.582–0.600)	0.203 (0.195–0.212)
CKLE	0.731 (0.724–0.737)	0.267 (0.258–0.277)	0.700 (0.693–0.707)	0.265 (0.255–0.275)	0.670 (0.661–0.680)	0.251 (0.242–0.263)
Synthetic Notes Augmentation	0.722 (0.716–0.728)	0.266 (0.256–0.276)	0.693 (0.685–0.700)	**0.275 (0.265–0.287)**	0.660 (0.651–0.669)	0.254 (0.244–0.265)
OC-Distill (Ours)	**0.746 (0.740–0.751)**	**0.274 (0.265–0.284)**	**0.703 (0.695–0.710)**	0.273 (0.263–0.285)	**0.675 (0.666–0.685)**	**0.273 (0.263–0.286)**

Table 16: External validation on eICU vital signs for in-hospital mortality prediction across 48, 72, and 96 hour observation horizons. Best results are **bold**.

Method	48h		72h		96h	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Transformer	0.640 (0.637–0.643)	**0.143 (0.141–0.144)**	0.625 (0.621–0.628)	0.140 (0.139–0.142)	0.608 (0.604–0.612)	**0.139 (0.137–0.141)**
Cross-Modal Contrastive	0.640 (0.637–0.643)	0.142 (0.141–0.144)	0.623 (0.619–0.626)	0.139 (0.138–0.141)	0.577 (0.573–0.581)	0.125 (0.124–0.127)
CKLE	**0.641 (0.638–0.644)**	0.142 (0.141–0.144)	0.624 (0.621–0.628)	0.140 (0.139–0.142)	**0.609 (0.606–0.613)**	0.138 (0.136–0.140)
Synthetic Notes Augmentation	0.639 (0.636–0.641)	0.142 (0.141–0.143)	0.620 (0.616–0.623)	0.139 (0.137–0.140)	0.606 (0.602–0.609)	0.137 (0.136–0.139)
OC-Distill (Ours)	0.639 (0.636–0.642)	0.141 (0.140–0.143)	**0.629 (0.626–0.633)**	**0.143 (0.141–0.145)**	0.609 (0.605–0.612)	0.138 (0.137–0.141)

Table 17: External validation on eICU vital signs for length-of-stay prediction across 48, 72, and 96 hour observation horizons. Best results are **bold**.