

Unveiling Discrete Clues: Superior Healthcare Predictions for Rare Diseases

Anonymous Author(s)

Abstract

Accurate healthcare prediction is essential for improving patient outcomes. Existing work primarily leverages advanced frameworks like attention or graph networks to capture the intricate collaborative (CO) signals in electronic health records. However, prediction for rare diseases remains challenging due to limited co-occurrence and inadequately tailored approaches. To address this issue, this paper proposes UDC, a novel method that unveils discrete clues to bridge consistent textual knowledge and CO signals within a unified semantic space, thereby enriching the representation semantics of rare diseases. Specifically, we focus on addressing two key sub-problems: (1) acquiring distinguishable discrete encodings for precise disease representation and (2) achieving semantic alignment between textual knowledge and the CO signals at the code level. For the first sub-problem, we refine the standard vector quantized process to include condition awareness. Additionally, we develop an advanced contrastive approach in the decoding stage, leveraging synthetic and mixed domain targets as hard negatives to enrich the perceptibility of the reconstructed representation for downstream tasks. For the second sub-problem, we introduce a novel codebook update strategy using co-teacher distillation. This approach facilitates bidirectional supervision between textual knowledge and CO signals, thereby aligning semantically equivalent information in a shared discrete latent space. Extensive experiments on three datasets showcase that the proposed UDC notably improves healthcare prediction performance for both rare and common diseases.

Keywords

Discrete modeling, Healthcare prediction, Rare disease

Relevance: *It uses web data for better patient modeling (user modeling), enhancing the Web's technical and socio-economic systems.*

1 Introduction

Healthcare predictions, such as medication recommendations, are critically important as they directly influence the efficacy of medical treatments [13, 30]. Accurate medication recommendations can enhance patient recovery rates by up to 30% and reduce adverse drug reactions by 25%, demonstrating their significant positive impact [37, 41]. However, erroneous predictions present a dual-edged sword: they can lead to ineffective treatments in approximately 20% of cases [7, 58]. This highlights the imperative to enhance the performance of healthcare predictions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '25, April 28 -May 02, 2025, Sydney, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/XXXXXXX.XXXXXXX>

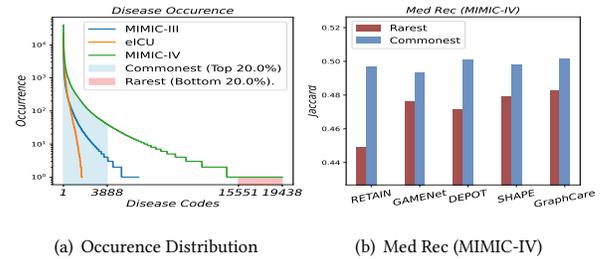


Figure 1: (a) Disease occurrences across three datasets. (b) Medication recommendation for commonest / rarest diseases.

Current research in healthcare prediction can be broadly categorized into three genres [1, 32, 49]: rule-based, graph-based, and sequence-based approaches. Rule-based systems [9, 38] typically rely on expert-defined rules to guide predictions, offering effective solutions but often facing limitations in scalability and potential conflicts among rules. In contrast, graph-based methods [3, 5] leverage graph neural networks to model electronic health records (EHRs) as homogeneous or heterogeneous graphs, enhancing predictive performance through the exploration of intricate collaborative (CO) signals within the data. Sequence-based methods [50, 56] represent a shift from static approaches by focusing on the sequential patterns inherent in longitudinal EHRs, capturing temporal dependencies that static models might overlook. While these methods are effective, they tend to emphasize maximizing overall accuracy [48, 58], which can lead to performance degradation for specific diseases. This issue arises from the highly skewed data distribution in EHRs. As depicted in Figure 1(a), datasets such as MIMIC-III [17], MIMIC-IV [16], and eICU [34] exhibit a pronounced imbalance in data distribution. In MIMIC-IV dataset, the commonest diseases (top 20%) account for approximately 95% of interactions in EHRs, while the rarest diseases (tail 20%) represent only about 0.2%. Meanwhile, as shown in Figure 1(b), we observe that existing advanced methods demonstrate superior performance in diagnosing common diseases. However, their effectiveness diminishes significantly when applied to rare diseases. This disparity is a key factor contributing to overall predictive shortcomings and may lead to health inequalities in diagnosis [58]. It underscores the need for more effective strategies.

Recently, several studies have demonstrated distinct distributions of long-tail and head objects [55]. This observation motivates us to treat rare diseases and common diseases as different feature domains and find a way to align rare diseases (CO space) with common diseases (CO space) to leverage the established knowledge, e.g., disease-medication relationships derived from rich EHRs associated with common diseases. However, as depicted in Figure 1, limited data impedes the establishment of a robust CO space for rare diseases. Textual knowledge (Text), shared across all diseases and recognized as a consistent and reliable semantic resource [18, 48], serves as a bridge to facilitate alignment between these two spaces. Consequently, our aim is to align CO signals with textual knowledge within a unified discrete space, followed by executing a high-quality

Text→CO mapping for rare diseases to enrich representation semantics. The discrete space, derived from VQ-VAE [36], employs a vector quantized (VQ) process to facilitate code-level mappings between textual knowledge and CO signals. This aligns with the multi-symptom nature of the disease and demands fewer computations compared to continuous modeling [10, 21]. To develop our approach, we highlight two key aspects.

- **How to acquire distinguishable discrete encodings for precise disease representation?** 1) In clinical documentation, even minor variations in symptoms can necessitate different medical codes, despite similar text descriptions. For instance, Type 1 and Type 2 diabetes, though both may present as "diabetes without complications," diverge significantly in their pathophysiology and management, with Type 1 typically requiring lifelong insulin therapy and Type 2 often managed through lifestyle modifications and oral medications. This necessitates that the model be adept at discerning subtle yet significant differences in clinical context, despite relatively similar text descriptions. 2) While VQ-VAE is effective at reconstructing data and learning broad patterns, its approach to feature extraction and reconstruction may not always align with the specific, detailed requirements of downstream predictive tasks, resulting in potential limitations in predictive accuracy. For example, while the reconstructed text representation provides a coherent overview, it might lack critical details like specific symptom patterns or treatment adherence levels. Similarly, reconstructed CO signals might miss key interactions or subtle patterns that are crucial for precise medication recommendation or diagnosis prediction.
- **How to perform effective semantic alignment between CO signals and textual knowledge?** Text and CO signals typically reside in distinct semantic spaces, with text represented in natural languages and CO signals in interaction embeddings. This domain gap is an obstacle that hinders the Text→CO signal mapping. Furthermore, as both representations of disease are mapped into a discrete space—where each code embodies unique symptom semantics—aligning at the code level is crucial for mitigating the domain gap and facilitating knowledge transfer.

To tackle these challenges, we introduce UDC, a tailored VQ-VAE framework for healthcare that utilizes textual knowledge and CO signals for alignment and reconstruction, enhancing the representation semantics of rare diseases during discrete representation learning (DRL). To ensure the distinguishability of disease encodings, we upgrade the original VQ process to incorporate condition-aware calibration. We specifically include medical entities that co-occur during the same visit for a particular disease as contextual conditions. This adjustment allows the model to produce distinct reconstructions based on varying contexts, even when the text appears similar. For instance, in a medical scenario, the distinction between Type 1 and Type 2 diabetes could be identified by examining complications such as diabetic ketoacidosis (more common in Type 1) or by specific laboratory findings in EHRs, thereby enhancing the granularity of representations. Furthermore, to guarantee task relevance in the reconstructed representations, we devise a contrastive task-aware calibration. Leveraging mixed-domain and synthetic target representations as hard negatives, we boost the model's ability to discern distinct features and facilitate the reciprocal transfer of knowledge between CO signals and textual information. This

empowers the reconstructed representations to react adaptively in accordance with the particular downstream tasks at hand. To achieve better semantic alignment of Text-CO signals, we introduce a novel codebook update strategy using co-teacher distillation. In this approach, the text and the CO signal, both featuring encoded diseases, act as mutual reconstruction labels, facilitating the aggregation of quantized vectors encoded from two signals with equivalent semantics into a unified latent space.

To sum up, our key contributions are as follows.

- To our knowledge, UDC has significantly enriched the semantics of rare diseases, thereby improving healthcare prediction performance. Our framework can be seamlessly integrated into various advanced healthcare prediction models.
- We tailor the VQ process for healthcare, incorporate condition-aware and task-aware calibration, and devise a novel codebook update mechanism. These enhancements notably improve reconstruction performance and adaptability to downstream tasks.
- Our algorithm demonstrates superior performance across two healthcare prediction tasks on three datasets, effectively handling both common and rare diseases. We have made the code available on Github ¹ to ensure reproducibility.

2 Related Work

We review related work, emphasizing connections and distinctions.

2.1 Healthcare Prediction

Healthcare prediction employs advanced data-driven models to forecast clinical outcomes and disease progression [49]. This practice significantly impacts personalized treatment by facilitating early intervention and optimizing clinical decisions.

The primary genres in healthcare prediction include rule-based, graph-based, and sequence-based models. Rule-based models [9, 38], stemming from clinical expertise, offer interpretability and ease of implementation. However, their limitations lie in adapting to dynamic patient data and conflict rules, hindering their efficacy. In contrast, graph-based models diverge as they are entirely data-driven. They intricately map relationships among clinical entities as nodes and edges within a graph framework [3, 5], excelling in modeling relational data and uncovering hidden patterns. However, they can be computationally intensive and encounter scalability challenges when applied to extensive datasets. On the other hand, sequence-based models [11, 26], leveraging temporal data like longitudinal EHRs, dynamically capture temporal dependencies. This paradigm is typically constructed using architectures such as RNNs and Transformers. When combined with medical prior knowledge, it effectively captures the patient's condition. Recently, hybrid models [14, 48] have been introduced, combining these genres to harness their respective strengths. A common approach involves representing visit-level data as subgraphs or introducing external knowledge, followed by information extraction to incorporate both temporal and high-order CO signals. While effective, most methods primarily aim to enhance overall accuracy, with limited focus on the unique challenges associated with the sparse rare diseases.

Our approach operates within this hybrid genre, specifically targeting the enhancement of rare disease prediction through the integration

¹<https://anonymous.4open.science/r/UDCHHealth/README.md>

of textual knowledge. Leveraging discrete learning, our method effectively bridges textual knowledge with CO signals, bolstering the representation semantics tailored to rare diseases.

2.2 Generative Retrieval

Generative retrieval is a key technique in modern systems, enabling the direct generation of candidate items rather than selecting from a fixed set, as in discriminative genres [20]. This is critical for delivering context-aware retrievals in domains with limited data.

Generative retrieval [22, 39] can be broadly categorized into three genres: autoregressive-based [24, 47], GAN-based [4, 15], and autoencoder-based models [2, 46, 59]. Autoregressive models [24], such as those utilizing Transformer architectures, generate sequences by predicting the next item based on previous context, making them well-suited for tasks requiring a sequential understanding. However, they are often computationally intensive and may suffer from exposure bias. GAN-based models [4] generate realistic candidate items through a generator that creates samples and a discriminator that evaluates their authenticity. While GANs [15] excel in producing high-quality outputs, they are challenging to train and may experience instability issues. Autoencoder-based models, including approaches like VAE [35, 54], use an encoder to map inputs to a latent space and a decoder to reconstruct them. These models effectively capture complex data distributions and facilitate structured, interpretable generation. VQ-VAE [36], in particular, leverages discrete latent variables, balancing the strengths of both autoregressive and autoencoder-based approaches while offering robustness in handling diverse distributions.

Our method aligns with the last genre, specifically extending VQ-VAE to healthcare. We focus on enhancing the representation of rare diseases by introducing condition-aware and task-aware calibration. Furthermore, we devise a novel co-teacher distillation to achieve code-level semantic alignment. These tailored advancements enhance the accuracy and relevance of the rare disease representations generated, thereby boosting the performance of VQ-VAE within healthcare tasks.

3 Proposed Method

Preliminary. Each patient’s medical history is recorded as a sequence of visits, represented by $\mathcal{U}^{(k)} = (\mathbf{u}_1^{(k)}, \mathbf{u}_2^{(k)}, \dots, \mathbf{u}_{\mathcal{T}_k}^{(k)})$, where k identifies the patient within the patient set \mathcal{N} , and \mathcal{T}_k is the total number of visits. Each visit $\mathbf{u}_t^{(k)}$ is defined as a triplet $\mathbf{u}_t^{(k)} = (\mathbf{d}_t^{(k)}, \mathbf{p}_t^{(k)}, \mathbf{m}_t^{(k)})$, corresponding to the diagnoses (d), procedures (p), and medications (m) associated with that visit, respectively. These components are encoded as multi-hot vectors: $\mathbf{d}_t^{(k)} \in \{0, 1\}^{|\mathcal{D}|}$, $\mathbf{p}_t^{(k)} \in \{0, 1\}^{|\mathcal{P}|}$, and $\mathbf{m}_t^{(k)} \in \{0, 1\}^{|\mathcal{M}|}$, where \mathcal{D} , \mathcal{P} , and \mathcal{M} represent the sets of all possible diagnoses, procedures, and medications, and $|\cdot|$ denotes the cardinality of these sets. For instance, the vector $\mathbf{d} = [1, 0, 1, 0]$ suggests that the patient has diseases 1 and 3, assuming $|\mathcal{D}| = 4$. Additionally, each medical entity $*$ is associated with a corresponding text description denoted as $T(*)$. For clarity, k is omitted in the following content.

Task formulation. Following [52, 53, 56], we outline the definitions of the two common healthcare prediction tasks.

- **Diagnosis Prediction (Diag Pred)** entails a multi-label classification challenge that centers on anticipating forthcoming risks.

This task revolves around scrutinizing $[\mathbf{u}_1, \dots, \mathbf{u}_t]$ to forecast the diagnosis set \mathbf{d}_{t+1} at time $t + 1$, where target $\mathbf{y}[\mathbf{u}_{t+1}] \in \mathbb{R}^{1 \times |\mathcal{D}|}$.

- **Medication Recommendation (Med Rec)** involves a multi-label classification task dedicated to pinpointing the most suitable medications for the patient’s present state. This process entails scrutinizing $[\mathbf{u}_1, \dots, \mathbf{u}_t]$, alongside $(\mathbf{d}_{t+1}, \mathbf{p}_{t+1})$, to anticipate \mathbf{m}_{t+1} at time $t + 1$, where target $\mathbf{y}[\mathbf{u}_{t+1}] \in \mathbb{R}^{1 \times |\mathcal{M}|}$.

Notations for important symbols could be found in Appendix A.

Solution Overview. Our solution for enhancing healthcare prediction, particularly for rare diseases, unfolds through a structured three-step process. First, we develop a robust healthcare prediction model $\mathcal{F}_{\text{co}}(\cdot)$ by training on the entire dataset, which acts as the pre-trained collaborative model (PCM). However, this alone proves insufficient, as the resulting representations $\mathbf{E}_{\mathcal{D}}$ often fail to capture the nuances of rare diseases due to sparse co-occurrence. To address this, we choose a pre-trained language model (PLM), i.e. $\mathcal{F}_{\text{te}}(\cdot)$, and introduce a discrete representation learning (DRL) framework in the second stage, where we reconstruct these representations to ensure Text-CO signals alignment. Our key innovations lie in this phase, where we employ condition injection, contrastive learning, and co-teacher distillation to ensure that the discretized representations, incorporating both textual and collaborative signals, are distinct, task-aware, and aligned at the code level. Finally, in the fine-tuning & inference stage, we freeze DRL to produce $\tilde{\mathbf{E}}_{\mathcal{D}}$ that substitute the original embeddings $\mathbf{E}_{\mathcal{D}}$ and fine-tune $\mathcal{F}_{\text{co}}(\cdot)$, thereby significantly improving the model’s capability to handle the challenging rare cases. The comprehensive framework is illustrated in Figure 2.

3.1 Discrete Disease Representation

We employ discrete modeling to map disease representations onto discretized code vectors for reconstruction. Contrasted with VAEs, VQ [21] process excels in compression and offers interpretability.

Pre-trained PCM & PLM. Initially, we train a conventional healthcare prediction model optimized with commonly used binary cross-entropy (BCE) [14, 43], employing EHRs to construct collaborative representations for each medical entity. Formally,

$$\mathbf{e}_d = \mathbf{E}_{\mathcal{D}}(d), \quad \mathbf{e}_p = \mathbf{E}_{\mathcal{P}}(p), \quad \mathbf{e}_m = \mathbf{E}_{\mathcal{M}}(m), \quad (1)$$

$$\mathcal{L}_{\text{task}} = \text{BCE}(\mathbf{y}, \mathcal{F}_{\text{co}}(\mathbf{e}_d, \mathbf{e}_p, \mathbf{e}_m, \mathcal{T}_k; \theta)), \quad (2)$$

where $\mathcal{F}_{\text{co}}(\cdot)$ can denote any PCM. Here, we opt for Transformer [40] as the backbone. As evidenced in [1, 44, 57], $\mathcal{F}_{\text{co}}(\cdot)$ extracts interaction patterns, whereas embedding \mathbf{E} encompasses rich CO similarities. Likewise, we choose a popular clinical pre-trained language model $\mathcal{F}_{\text{te}}(\cdot)$, i.e. Sap-BERT [25], to serve as the PLM. Formally,

$$\tilde{\mathbf{e}}_d = \tilde{\mathbf{E}}_{\mathcal{D}}(T(d)), \quad \tilde{\mathbf{e}}_p = \tilde{\mathbf{E}}_{\mathcal{P}}(T(p)), \quad \tilde{\mathbf{e}}_m = \tilde{\mathbf{E}}_{\mathcal{M}}(T(m)), \quad (3)$$

where $\tilde{\mathbf{E}}$ signifies the embedding table of $\mathcal{F}_{\text{te}}(\cdot)$. We contrast the variations among various PCM and PLM backbones in Section 4.3.4.

Discrete Representation. Next, we consider mapping the disease encoding \mathbf{e}_d and $\tilde{\mathbf{e}}_d$ to a set of discrete codes using RQ-VAE [21], a widely adopted VQ-VAE framework. In RQ-VAE, L-level codebooks are defined. For each code-level $l \in \{1, \dots, L\}$, there exists a codebook $C_l = \{\mathbf{c}_i\}^{C_l}$. Subsequently, for disease d , the associated set of discrete codes is derived through the residual method. Formally,

$$\begin{cases} c_l = \arg \min_i \|\mathbf{r}_{l-1} - \mathbf{c}_i\|_2, & \mathbf{c}_i \in C_l, \\ \mathbf{r}_l = \mathbf{r}_{l-1} - \mathbf{c}_l, \end{cases} \quad (4)$$

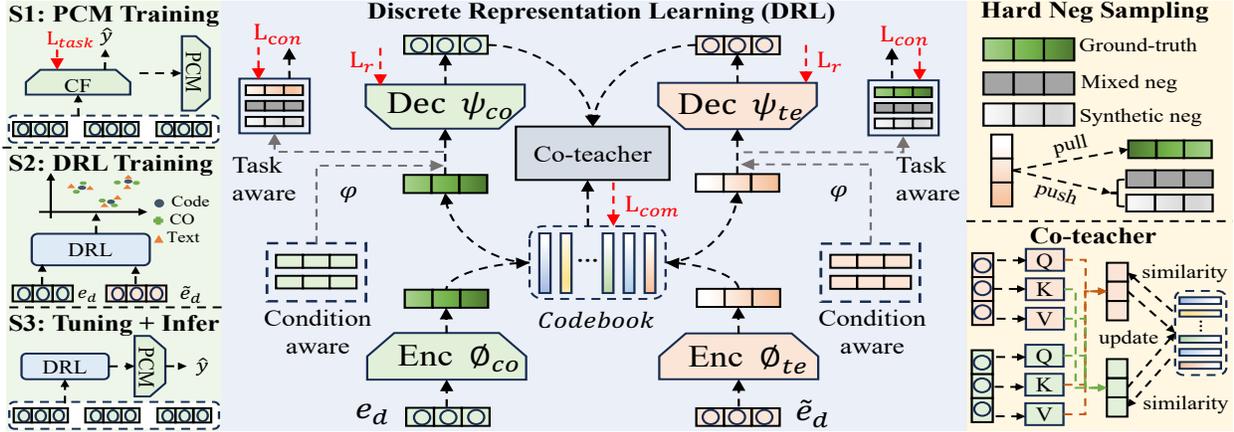


Figure 2: Overview of UDC. We pre-train the PCM to establish a robust CO space and then obtain CO and text representations for diseases using PCM and a selected PLM. Next, we train the DRL to align the text and CO signals, followed by fine-tuning the PCM for downstream tasks while keeping the DRL frozen. Q, K, V denotes the parameters for multi-head attention.

where c_l denotes the assigned code index from the l -th level codebook and $\|\cdot\|_2$ is 2-Norm. r_{l-1} is the semantic residual from the last level and we set $r_0 = \phi_{co}(e_d)$ or $\tilde{r}_0 = \phi_{te}(\tilde{e}_d)$, where ϕ is an MLP encoder layer. Finally, for each medical entity, we have the discrete PCM codes and discrete PLM codes, i.e., $e_d \rightarrow \{c_1, c_2, \dots, c_L\}$, $\tilde{e}_d \rightarrow \{\tilde{c}_1, \tilde{c}_2, \dots, \tilde{c}_L\}$. For efficiency, we utilize a shared codebook for both text and CO signals, i.e., $\tilde{c}_l \in C_l$. Then we get the encoded disease representation using the sum operation. Formally,

$$\mathbf{z}_d = \sum_{l=1}^L c_l, \quad \tilde{\mathbf{z}}_d = \sum_{l=1}^L \tilde{c}_l, \quad (5)$$

where \mathbf{z}_d and $\tilde{\mathbf{z}}_d$ denote the discrete representation for a disease. In other words, we discretize the disease into the sum of various symptom codes, offering a more intuitive approach.

3.2 Condition-aware Calibration

Traditional RQ-VAE typically proceeds to decode once the latent vector \mathbf{z}_d is obtained. However, their efficacy in reconstructing samples with similar descriptions is limited. This constraint significantly hampers their utility in healthcare scenes, where medical entities frequently share analogous descriptions yet possess distinct semantic nuances. To address this deficiency, we propose integrating external conditions, specifically diverse types of medical entities within the same visit, to modulate the quantization vector via normalization. This strategy aims to embed condition variations into the index map, thereby stimulating the decoder to produce a broader array of reconstructed representations. Formally,

$$\mathbf{f}_d = \text{MHA}_{\mathcal{P}}(\mathbf{e}_p^d, \mathbf{e}_p^d, \mathbf{e}_p^d) + \text{MHA}_{\mathcal{M}}(\mathbf{e}_m^d, \mathbf{e}_m^d, \mathbf{e}_m^d), \quad (6)$$

where $\text{MHA}(\cdot)$ denotes multi-head attention and \mathbf{f}_d refers to the condition representation. $\mathbf{e}_p^d \in \mathcal{E}_{\mathcal{P}}$ and $\mathbf{e}_m^d \in \mathcal{E}_{\mathcal{M}}$ refer to the entities corresponding to disease d at the same visit. Then, we incorporate it in normalized form. Formally, for the CO branch,

$$\mathbf{z}_d = \varphi_{\gamma}(\mathbf{z}_d^{\text{old}}) \frac{\mathbf{f}_d - \mu(\mathbf{f}_d)}{\sigma(\mathbf{f}_d)} + \varphi_{\beta}(\mathbf{z}_d^{\text{old}}), \quad (7)$$

where $\mathbf{z}_d^{\text{old}}$, as defined in Eq. 5, is labeled as "old" for clarity. μ and σ denotes the mean and variation. φ_{γ} and φ_{β} signify the transformation matrix. This normalizing ensures that \mathbf{f} 's values fall within

a similar range, which helps maintain consistency in the scale of the input features, thereby aiding in training stability and convergence without escalating the model's complexity. Likewise, we could obtain $\tilde{\mathbf{z}}_d$ using \tilde{e}_p^d and \tilde{e}_m^d .

3.3 Task-aware Calibration

While incorporating conditions can enhance the semantics of \mathbf{z}_d for decoding, there remains a crucial gap: the model lacks awareness of downstream tasks. This awareness can help optimize model performance by guiding the learning process towards features that are most relevant to the healthcare task, leading to improved accuracy. In other words, we necessitate that the reconstructed representation not only mirrors the original one but also closely aligns with the target \mathcal{S}_d in the subsequent visit ($\mathcal{S}_d \in \mathcal{D}$ for Diag Pred and $\mathcal{S}_d \in \mathcal{M}$ for Med Rec); otherwise, it remains distant. To achieve this objective, beyond conventional intra-domain (Text/CO signal) contrastive learning [27], we devise two distinct hard negative sampling to augment the contrastive training approach. Formally, using CO signal \mathbf{z}_d as an example,

$$\mathcal{L}_{\text{intra}} = -\frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} \log \left[\frac{\exp(\mathbf{s}_d W \mathbf{z}_d)}{\underbrace{\exp(\mathbf{s}_{d'} W \mathbf{z}_d)}_{\text{synthetic}} + \sum_{j \neq d} \underbrace{\exp(\mathbf{s}_j W \mathbf{z}_d)}_{\text{intra-domain}}} \right], \quad (8)$$

$$\mathcal{L}_{\text{inter}} = -\frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} \log \left[\frac{\exp(\tilde{\mathbf{s}}_d W \mathbf{z}_d)}{\underbrace{\exp(\tilde{\mathbf{s}}_{d'} W \mathbf{z}_d)}_{\text{synthetic}} + \sum_{j \neq d} \underbrace{\exp(\tilde{\mathbf{s}}_j W \mathbf{z}_d)}_{\text{mixed domain}}} \right], \quad (9)$$

where \mathbf{s}_d denotes d 's next-visit target representation, i.e., $\mathbf{s}_d = \sum_{d' \in \mathcal{S}_d} \phi_{co}(\mathbf{e}_{d'})$. $\mathbf{s}_{d'}$ denotes the synthetic disease representation acquired by randomly substituting the medical entities associated with the target \mathcal{S}_d . Likewise, we define $\tilde{\mathbf{s}}_d = \sum_{d' \in \mathcal{S}_d} \phi_{te}(\tilde{\mathbf{e}}_{d'})$. Formally, we advance from both collaborative and textual standpoints,

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}} + \tilde{\mathcal{L}}_{\text{intra}} + \tilde{\mathcal{L}}_{\text{inter}}, \quad (10)$$

where $\tilde{\mathcal{L}}$ signifies the contrastive learning using $\tilde{\mathbf{z}}_d$. This bidirectional learning ensures that the representations reconstructed by PCM and PLM not only encapsulate the relevance within the domain but also encompass the similarity of entities across domains.

3.4 Co-teacher Distillation

In the preceding sections, we transform both the CO and textual signals into discrete representations. However, this pipeline does not ensure semantic alignment between the two at the code level, leading to a domain gap that significantly impedes the subsequent Text→CO mapping. To address this constraint, we introduce a co-teacher distillation that iteratively refines the same code by leveraging both text and CO signals. Specifically, for each code c_i , we first retrieve the related diseases set N_i^l and \tilde{N}_i^l in the collaborative and textual domain at the l -th level codebook. Subsequently, we combine their representations to obtain a holistic view \mathbf{o}^l . For clarity, we omit the superscript l . Formally, for t -th iteration,

$$\begin{aligned} \mathbf{o}_i^{(t)} &= \kappa \mathbf{o}_i^{(t-1)} + (1 - \kappa) \left[\sum_{d \in N_i^{(t)}} \frac{\mathbf{z}_d^{(t)} + \tilde{\mathbf{b}}_d^{(t)}}{2} + \sum_{d \in \tilde{N}_i^{(t)}} \frac{\tilde{\mathbf{z}}_d^{(t)} + \mathbf{b}_d^{(t)}}{2} \right], \\ \mathbf{b}_d^{(t)} &= \text{MHA}(\mathbf{z}_d, \tilde{\mathbf{z}}_d, \tilde{\mathbf{z}}_d), \quad \tilde{\mathbf{b}}_d^{(t)} = \text{MHA}(\tilde{\mathbf{z}}_d, \mathbf{z}_d, \mathbf{z}_d), \end{aligned} \quad (11)$$

where κ refers to the decay rate and \mathbf{b} extract the relationship between two views. Then, we employ an exponential moving average method to update c_i . Formally,

$$\begin{aligned} \mathbf{c}_i^{(t)} &= \mathbf{o}_i^{(t)} / \mathbf{n}_i^{(t)}, \\ \mathbf{n}_i^{(t)} &= \kappa \mathbf{n}_i^{(t-1)} + (1 - \kappa) \left[\sum_{d \in N_i^{(t)}} \mathbf{z}_d^{(t)} + \sum_{d \in \tilde{N}_i^{(t)}} \tilde{\mathbf{z}}_d^{(t)} \right], \end{aligned} \quad (12)$$

where \mathbf{n}_i are used for normalization. We also modify the commitment loss in RQ-VAE by utilizing the code vector $\tilde{\mathbf{z}}_d$ as a teacher to guide the encoder ϕ_{co} . This modification aims for $\phi_{\text{co}}(\mathbf{e}_d)$ to not only approximate \mathbf{z}_d but also to converge towards $\tilde{\mathbf{z}}_d$ at a ratio of 50% in our setting, with α is the commitment weight. Formally,

$$\begin{aligned} \mathcal{L}_{\text{com}} &= \underbrace{\alpha \|\phi_{\text{co}}(\mathbf{e}_d) - \text{sg}[\mathbf{z}_d]\|_2^2}_{\text{origin}} + \underbrace{\frac{\alpha}{2} \|\phi_{\text{co}}(\mathbf{e}_d) - \text{sg}[\tilde{\mathbf{z}}_d]\|_2^2}_{\text{new}} \\ &+ \underbrace{\alpha \|\phi_{\text{te}}(\tilde{\mathbf{e}}_d) - \text{sg}[\tilde{\mathbf{z}}_d]\|_2^2}_{\text{origin}} + \underbrace{\frac{\alpha}{2} \|\phi_{\text{te}}(\tilde{\mathbf{e}}_d) - \text{sg}[\mathbf{z}_d]\|_2^2}_{\text{new}}, \end{aligned} \quad (13)$$

where sg denotes the stop gradient. This alignment compels the CO signal and the textual space to converge on the same symptom code at each discrete level and maintain the consistent code semantics, thereby facilitating subsequent representation substitution.

3.5 Training & Fine-tuning Strategy

We outline the training objectives of the DRL and fine-tuning stages. **Training Strategy.** Our final optimization objective for DRL comprises reconstruction loss and the two preceding parts. Formally,

$$\mathcal{L}_{\text{total}} = \underbrace{\|\mathbf{e}_d - \psi_{\text{co}}(\mathbf{z}_d)\|_2^2 + \|\tilde{\mathbf{e}}_d - \psi_{\text{te}}(\tilde{\mathbf{z}}_d)\|_2^2}_{\text{reconstruction loss } \mathcal{L}_r} + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{com}}, \quad (14)$$

where ψ denotes the MLP decoder for reconstruction. Once DRL is trained, it can be used as a mapping function to transform textual space into collaborative space. At this stage, we exclusively leverage data related to common diseases \mathcal{D}_{com} , as collaborative signals from rare diseases \mathcal{D}_{rar} are considered unreliable. \mathcal{D}_{com} and \mathcal{D}_{rar} are splited according to Section 4.1.

Fine-tuning & Inference. Upon DRL alignment training completion, DRL can transform textual signals into collaborative signals.

This enables us to utilize the textual description of rare diseases to supplant their original inferior collaborative signals. Formally,

$$\hat{\mathbf{e}}_d = \begin{cases} \psi_{\text{co}}[\varphi(\phi_{\text{te}}(\tilde{\mathbf{e}}_d); \mathbf{e}_p^d, \mathbf{e}_m^d)], & \text{if } d \in \mathcal{D}_{\text{rar}} \\ \psi_{\text{co}}[\varphi(\phi_{\text{co}}(\mathbf{e}_d); \mathbf{e}_p^d, \mathbf{e}_m^d)], & \text{if } d \in \mathcal{D}_{\text{com}} \end{cases}. \quad (15)$$

Following this, we freeze DRL and $\mathbf{E}_{\mathcal{D}}$, and fine-tune $\mathcal{F}_{\text{co}}(\cdot)$ to capture updated interaction patterns using Eq. 2. This step is crucial, as evidenced in Appendix B, since the prior $\mathcal{F}_{\text{co}}(\cdot)$ may not fully grasp interaction patterns with other medical entities owing to the data scarcity on rare diseases. For a new representation, it necessitates re-learning to enhance its effectiveness. Finally, we can integrate $\mathcal{F}_{\text{co}}(\cdot)$ and DRL for the estimation $\hat{\mathbf{y}}$. Formally,

$$\hat{\mathbf{y}} = \mathcal{F}_{\text{co}}(\hat{\mathbf{e}}_d, \mathbf{e}_p, \mathbf{e}_m, \mathcal{T}_k; \theta). \quad (16)$$

Overall, through the three-step process, we can effectively map rare diseases onto the feature space of common diseases using textual knowledge as a bridge, thereby enhancing their semantic richness. A concise algorithm flow can be seen in Appendix C.

4 Experiments

We first outline the necessary setup and then present the analysis.

4.1 Experimental Setup

Datasets & Baselines. Our experiments are conducted on three popular healthcare datasets: MIMIC-III [17], MIMIC-IV [16], and eICU [34]. Detailed statistics for these datasets are summarized in Appendix D. Textual knowledge is extracted by parsing EHR entities according to the internationally recognized ICD and ATC systems [12] to obtain corresponding textual descriptions. We retain patients with more than one visit in MIMIC-III and eICU, while for MIMIC-IV, we include patients with two or more visits.

We select advanced baselines for comparison. Specifically, for both tasks, we include Transformer [40], MICRON [45], RETAIN [6], GRAM [5], StageNet [11], SHAPE [26], StratMed [23], HAR [43], GraphCare [14], and RAREMed [58]. For Diag Pred, we further incorporate HITANet [28], Deepr [33], Dipole [31], MedPath [53], SeqCare [48] as specialized baselines. In Med Rec, additional baselines such as SafeDrug [51], GAMENet [38], COGNet [50], VITA [19], MoleRec [52], and DEPOT [56], are included, given their distinctive designs and strong performance. Transformer, RETAIN, HITANet, Deepr, StageNet, RAREMed, and SHAPE are sequence-based approaches, while GRAM, GAMENet, MoleRec, MICRON, DEPOT, StratMed, COGNet, and VITA further integrate EHR graphs to enhance representation. MedPath, HAR, SeqCare, and GraphCare leverage external knowledge to improve performance. RAREMed and SeqCare incorporate tailored reconstruction tasks and denoising techniques specifically designed for rare diseases.

Implementation Details & Evaluations. To ensure fairness, all algorithms use an embedding dimension of 128. We employ the AdamW optimizer with a learning rate of 1e-3 for Diag Pred and 2e-4 for Med Rec. The batch size is set to 16. The epochs for the DRL and fine-tuning stages are set at 50 and 50, respectively. Following RQ-VAE, we configure the code layer $L = 4$, meaning each disease is represented by four codes. The codebook size $|C_l|$ and commitment weight α , which are crucial hyperparameters, are set to 64 and 0.25, respectively. Their effects are evaluated in Appendix F. Following the Pareto principle and previous research [55], we classify diseases appearing in 20% or more cases as common \mathcal{D}_{com} , with all others

Table 1: Performance comparison: Diagnosis Prediction. K=20.

Dataset	MIMIC-III				MIMIC-IV				eICU			
	Method	Acc@K	Pres@K	AUPRC	AUROC	Acc@K	Pres@K	AUPRC	AUROC	Acc@K	Pres@K	AUPRC
Transformer	0.2841	0.3144	0.2289	0.9174	0.3047	0.3420	0.2476	0.9591	0.6431	0.7716	0.6777	0.9667
MICRON	0.2735	0.3025	0.2130	0.9147	0.3081	0.3434	0.2098	0.9545	0.6308	0.7748	0.6781	0.9698
DeepR	0.2834	0.3132	0.2277	0.9113	0.2615	0.2904	0.1998	0.9396	0.6304	0.7620	0.6430	0.9584
HITANet	0.2917	0.3228	0.2309	0.9180	0.2996	0.3368	0.2432	0.9574	0.6517	0.7767	0.6773	0.9644
RETAIN	0.2920	0.3284	0.2509	0.9175	0.3078	0.3314	0.2337	0.9427	0.6576	0.7805	0.6879	0.9613
GRAM	0.3190	0.3559	0.2631	0.9182	0.3024	0.3513	0.2318	0.9591	0.6452	0.7891	0.6993	0.9711
Dipole	0.3183	0.3587	0.2631	0.9158	0.2968	0.3336	0.2395	0.9593	0.6585	0.7864	0.6677	0.9643
StageNet	0.3011	0.3375	0.2408	0.9188	0.3153	0.3440	0.2489	0.9593	0.6599	0.7936	0.6645	0.9702
SHAPE	0.3214	0.3531	0.2593	0.9226	0.3170	0.3540	0.2407	0.9564	0.6510	0.7779	0.6850	0.9676
StratMed	0.3076	0.3425	0.2434	0.9225	0.3137	0.3602	0.2595	0.9531	0.6449	0.7663	0.6710	0.9653
MedPath	0.3189	0.3490	0.2560	0.9224	0.3203	0.3616	0.2589	0.9620	0.6600	0.7947	0.7000	0.9714
HAR	0.3204	0.3532	0.2599	0.9193	0.3224	0.3642	0.2605	0.9628	0.6540	0.7910	0.6995	0.9720
GraphCare	0.3213	0.3529	0.2595	0.9203	0.3220	0.3635	0.2593	0.9620	0.6569	0.7788	0.6795	0.9694
SeqCare	0.3245	0.3547	0.2616	0.9213	0.3233	0.3668	0.2669	0.9632	0.6639	0.7996	0.7043	0.9727
RAREMed	0.3208	0.3521	0.2596	0.9192	0.3153	0.3527	0.2390	0.9557	0.6572	0.7792	0.6768	0.9692
<i>UDC</i>	0.3377	0.3713	0.2737	0.9256	0.3324	0.3707	0.2735	0.9657	0.6724	0.8070	0.7140	0.9736

Table 2: Performance comparison: Medication Recommendation.

Dataset	MIMIC-III				MIMIC-IV				eICU			
	Method	Jaccard	F1-score	AUPRC	AUROC	Jaccard	F1-score	AUPRC	AUROC	Jaccard	F1-score	AUPRC
Transformer	0.5012	0.6556	0.7671	0.9440	0.4635	0.6203	0.7305	0.9402	0.1159	0.3504	0.3138	0.9147
MICRON	0.4937	0.6501	0.7651	0.9307	0.4608	0.6123	0.7283	0.9362	0.0703	0.2349	0.2561	0.9017
SafeDrug	0.4859	0.6403	0.7367	0.9331	0.4569	0.6086	0.7293	0.9378	0.1061	0.4274	0.3036	0.9181
RETAIN	0.5049	0.6601	0.7680	0.9448	0.4646	0.6174	0.7364	0.9414	0.1181	0.4736	0.2835	0.9064
GRAM	0.4994	0.6537	0.7607	0.9435	0.4624	0.6155	0.7385	0.9424	0.0983	0.3166	0.2908	0.9168
GAMENet	0.5074	0.6612	0.7724	0.9456	0.4655	0.6181	0.7399	0.9425	0.1093	0.4165	0.2936	0.9103
COGNet	0.5114	0.6614	0.7774	0.9470	0.4612	0.6125	0.7271	0.9356	0.1166	0.3528	0.3237	0.9147
StageNet	0.5013	0.6494	0.7519	0.9358	0.4679	0.6201	0.7404	0.9424	0.1337	0.2303	0.3075	0.9201
VITA	0.5146	0.6671	0.7781	0.9469	0.4715	0.6219	0.7486	0.9424	0.1218	0.3640	0.3223	0.9157
MoleRec	0.5080	0.6624	0.7719	0.9451	0.4720	0.6254	0.7473	0.9411	0.1123	0.3609	0.3280	0.9219
DEPOT	0.5135	0.6697	0.7745	0.9466	0.4780	0.6298	0.7534	0.9465	0.1367	0.3875	0.3276	0.9134
SHAPE	0.5155	0.6678	0.7788	0.9469	0.4830	0.6347	0.7486	0.9475	0.1338	0.4056	0.3123	0.9154
StratMed	0.5070	0.6612	0.7724	0.9456	0.4719	0.6249	0.7446	0.9446	0.1223	0.3791	0.3031	0.9138
HAR	0.5126	0.6652	0.7758	0.9465	0.4805	0.6311	0.7539	0.9475	0.1257	0.4595	0.3153	0.9140
GraphCare	0.5167	0.6700	0.7805	0.9471	0.4816	0.6363	0.7576	0.9486	0.1252	0.4534	0.3107	0.9162
RAREMed	0.5134	0.6653	0.7786	0.9453	0.4794	0.6317	0.7496	0.9443	0.1304	0.4315	0.3119	0.9156
<i>UDC</i>	0.5261	0.6761	0.7833	0.9483	0.4912	0.6404	0.7580	0.9486	0.1443	0.4986	0.3296	0.9227

considered rare \mathcal{D}_{rar} . The impact of varying thresholds η is further explored in Appendix F.

For data partitioning, we follow established practices [43, 48, 56] by dividing the datasets into training, validation, and test sets in a 6:2:2 ratio. For Diag Pred, we use Acc@K, Pres@K, AUPRC, and AUROC for evaluation. Here K=20, different values are discussed in Section 4.3.2. For Med Rec, we assess using Jaccard, F1-score, PRAUC, and AUROC. These metrics are selected for their significant clinical relevance and comprehensive assessment [1, 14]. Their mathematical definitions are outlined in Appendix E.

Table 3: Ablation study. UDC-NCO does not incorporate condition-aware calibration. UDC-NT removes task-aware calibration. UDC-NM only leverages synthetic negative sampling. UDC-NS only utilizes mixed-domain negative sampling. UDC-NCD performs updates similar to RQ-VAE without using co-teacher distillation.

Algorithms	Metric	-NCO	-NT	-NM	-NS	-NCD	UDC
Diag Pred	Acc@K	0.3276	0.3297	0.3301	0.3318	0.3288	0.3377
	Pres@K	0.3606	0.3600	0.3608	0.3620	0.3591	0.3713
Med Rec	Jaccard	0.5176	0.5179	0.5205	0.5183	0.5171	0.5261
	F1-score	0.6703	0.6705	0.6709	0.6706	0.6689	0.6761

4.2 Overall Performance

As depicted in Tables 1-2, our proposed UDC achieves the best performance across all scenarios, despite only utilizing the relatively weak Transformer as the PCM. Regarding the baselines, we observe that the sequence-based methods, such as SHAPE and DEPOT significantly outperform GRAM, underscoring the importance of capturing temporal patterns. COGNet and VITA are Transformer variants that leverage medical priors, like EHR graphs, resulting in notable enhancements over pure Transformer. GraphCare, MedPath, and SeqCare distinguish themselves by leveraging external knowledge graphs to enrich the inherent entity semantics. Nevertheless, the absence of adequate denoising measures hinders their effectiveness. While RAREMed introduces pre-trained tasks to address the cold-start issue, its overall predictive capacity remains relatively modest. Observations suggest a potential decline for common disease prediction, as detailed in Section 4.3.3.

Concerning the tasks, Diag Pred is more challenging than Med Rec, as the former requires recalling and ranking a broader range of medical entities. UDC, GraphCare, and SeqCare demonstrate greater robustness, as they not only rely on CO signals but also leverage semantic associations between items from the external

knowledge. The broader Diag Pred benefits more from the external knowledge effects in the sampling process, leading to a 3% Acc@K improvement in MIMIC-IV. Our observations indicate that eICU demonstrates enhanced performance in Diag Pred, likely due to the smaller disease size, which results in greater similarity among diseases across consecutive periods. MICRON’s performance on MIMIC-III and eICU is constrained in both tasks due to its requirement for at least two visit lengths, which limits the available data. StratMed does not reproduce its success from Med Rec on Diag Pred. This disparity could stem from the drug interaction graph it introduced not being suitable for the Diag Pred.

Considering the datasets, MIMIC-IV is the most challenging, as it exhibits more complex entity interactions, reflected in the larger data volume and higher sparsity. Additionally, the MIMIC-IV data presents a more imbalanced distribution, as shown in Figure 1. Most algorithms, such as StratMed, Dipole, and DEPOT, experience noticeable performance degradation on this dataset. Despite incorporating external knowledge, as seen in GraphCare and HAR, their approaches overlook the domain gap between this knowledge and the CO signal, potentially leading to negative transfer. Meanwhile, the lack of standard EHR coding in the eICU dataset leads to significant gaps in external knowledge, diminishing the advantages of these baselines. Conversely, UDC directly leverages the text of eICU records and aligns CO signals with textual knowledge without requiring additional indexing, effectively alleviating this issue.

4.3 Model Analysis and Robust Testings

We conduct various robustness experiments on MIMIC-III to validate our efficacy. For Diag Pred, we choose SHAPE, RAREMed, and SeqCare, while for Med Rec, we additionally select GraphCare.

4.3.1 Ablation Study. We conduct ablation experiments to validate the efficacy of sub-modules. As shown in Table 3, UDC-CO, which lacks the condition-aware modeling between the disease and visit components, is the limited-effective configuration, with a substantial 3% drop in Diag Pred. This absence causes disease, akin to textual descriptions, to be challenging for the model to differentiate, thereby leading to a blurred decision boundary. While UDC-NT has little impact on the reconstruction ability, it fails to impose effective constraints on the representation space. Directly applying this representation to downstream tasks proves challenging, necessitating additional training during the fine-tuning phase, yet achieving equivalent performance remains elusive. When contrasted with UDC-NT, both UDC-NM and UDC-NS exhibit enhanced performance, attributed to their capability to enhance the model’s individual discernment by integrating hard negative instances. UDC-NCD, akin to RQ-VAE in codebook update, experiences a 2% degradation due to domain gaps between text and CO spaces. This disparity could result in a significant negative transfer. Overall, the results validate the essential contributions of the key sub-modules.

4.3.2 Examination of Top-K. Top-K evaluation is crucial as it strikes a balance between precise diagnosis and broad screening in Diag Pred [43, 48]. As shown in Figure 3, all algorithms’ Acc@K improve with increasing K, as a larger Top-K captures more relevant medical entities, aiding in the challenging Diag Pred. Notably, regardless of the specific K value setting, our UDC consistently outperforms the strongest baseline. Moreover, when $K = 10$, UDC demonstrates 4%

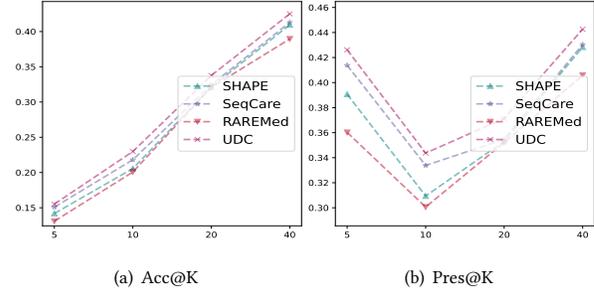


Figure 3: Top-K examination. We test $K = [5, 10, 20, 40]$.

improvement over the best competing SeqCare. This highlights the effectiveness of our approach in challenging scenarios. This further demonstrates its broad applicability, a crucial trait for clinical decision support systems, which often require flexibility in the number of diagnoses or treatment options presented.

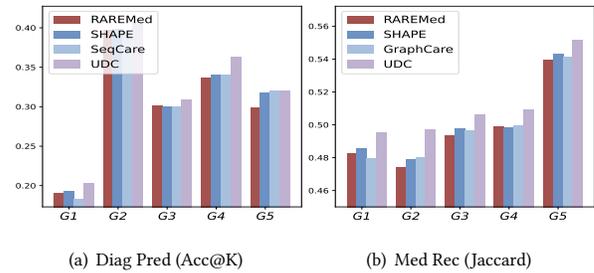


Figure 4: Group Analysis.

4.3.3 Group Analysis. To examine the model’s performance on rare diseases, we conduct a group-level analysis. Specifically, in Diag Pred, diseases are categorized into five prevalence groups: 0-20% (G1), 20-40%(G2), 40-60%(G3), 60-80%(G4), and 80-100%(G5), where G1 is the rarest disease group. As shown in Figure 4(a), the model’s efficacy in Diag Pred generally exhibits a positive correlation with the sparsity of the disease groups, with G2-G5 significantly outperforming G1. However, the performance of the G5 is not optimal, likely due to the low clinical significance of high-frequency diseases in Diag Pred; for instance, fever can indicate multiple underlying health risks. While RAREMed surpasses other baselines in G1 and G3, it compromises accuracy for common diseases. UDC exhibits the most notable boost in G1-G4, showcasing that our innovations excel at enhancing performance for rare diseases.

For the Med Rec, we further analyze the predictive performance for patient groups with various rare diseases. More precisely, we identify the rarest disease for each patient and allocate them to the corresponding group based on that rarity. Figure 4(b) indicates that recommendation performance for G1-G3 is limited, as fewer medications co-occur with their disease entities, leading to weaker disease-medication CO signals. Both SHAPE and RAREMed suffer from this issue. While GraphCare attempts to mitigate this problem by leveraging external knowledge, it fails to fully bridge the domain gap during the knowledge fusion and suffers from the potential knowledge noise. In contrast, UDC explicitly optimizes code-level alignment in DRL, facilitating bidirectional alignment of CO signals and textual knowledge, which leads to remarkable improvements.

In general, group-level analyses confirm that UDC significantly outperforms other baselines in managing rare diseases, essential for effective clinical decision support.

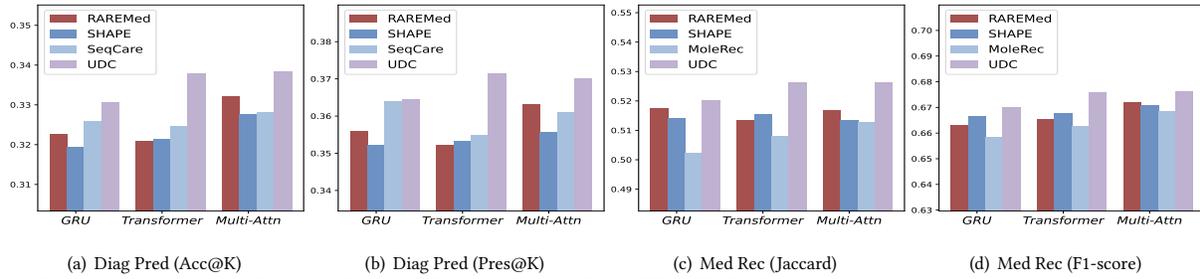


Figure 5: Plug-in Application (Diverse PCM). We choose MoleRec, SHAPE, RAREMed, and SeqCare, as they are flexible to PCM.

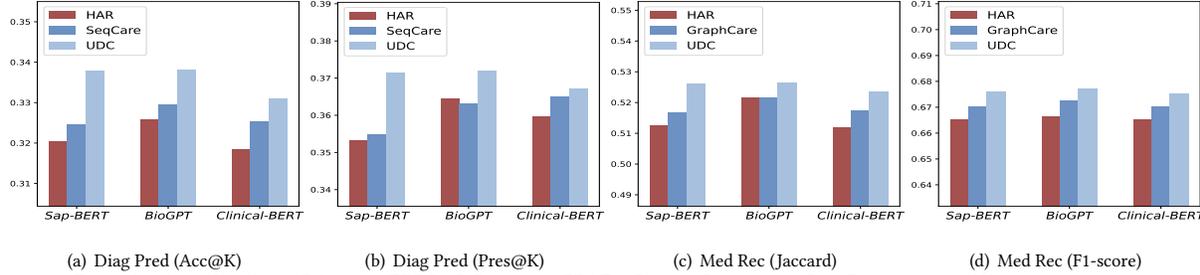


Figure 6: Plug-in Application (Diverse PLM). We select HAR, GraphCare, and SeqCare that utilize external knowledge.

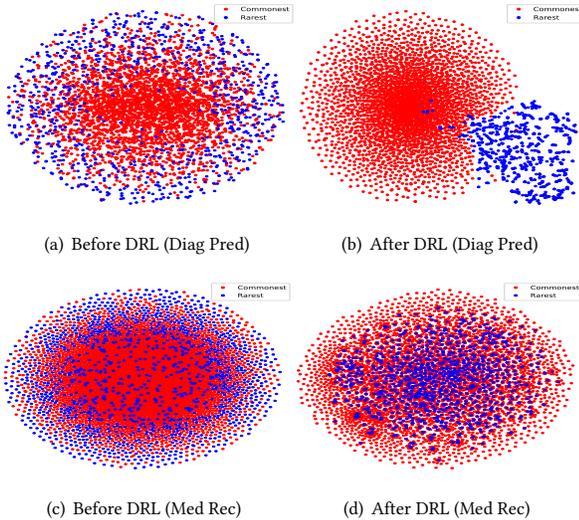


Figure 7: Code Semantics.

4.3.4 Plug-in Application. We examine the extensibility of UDC. **Diverse PCM.** For the PCM, we select three modern methods—GRU, Transformer, and Multi-head Attention—due to their widespread use in sequence-based healthcare baselines [31, 38, 52]. As shown in Figure 5, RAREMed has larger fluctuations, likely due to its explorations of three CO signals, maximizing its advantage from PCM. UDC demonstrates robust performance with various sophisticated PCM. The improvement in Multi-head Attention variants results from their significant CO advancements and convergence toward a more precise subspace during DRL alignment. This superior convergence contributes to an overall boost in model performance. **Diverse PLM.** Similarly, for the PLM, we evaluate the integration of both BioGPT [29] and Clinical-BERT [42]. Understanding the textual semantics encoded in clinical notes is another crucial aspect of the DRL, as it can capture similarities between entities that may not be evident from the EHRs alone. Compared to the Sap-BERT and Clinical-BERT, the BioGPT, which is fine-tuned on larger

medical-domain corpora, possesses more semantic representations. Furthermore, the larger parameter capacity of BioGPT enables it to obtain an even more robust alignment of the DRL module, leading to notable performance gains when integrated into UDC.

4.3.5 Case Study. We visualize disease representations before and after DRL. As shown in Figure 7, the rare disease prior to DRL exhibits a more random distribution with high entropy, indicating the PCM’s struggle to capture their inherent similarities. Because limited EHRs for rare diseases hinder the development of effective CO representations. In stark contrast, after DRL mapping, rare disease clusters are tightly grouped. This indicates that DRL effectively leverages text knowledge to capture underlying similarities, yielding more meaningful and clinically relevant representations of rare diseases. Another notable observation is that the distribution difference between rare and common diseases is more pronounced in Diag Pred than in Med Rec. This is because Diag Pred involves more complex relationships due to a higher number of targets, requiring greater changes in DRL. In contrast, Med Rec, with fewer targets, relies more on clearer existing models. This is intuitive, as medications are typically tailored for specific diseases and are relatively straightforward, whereas disease risks are often unpredictable.

5 Conclusion

In this paper, we introduce UDC, an innovative framework aimed at enhancing the representation semantics of rare diseases. UDC utilizes discrete representation learning to connect textual knowledge and CO signals, enabling both signals to be in the same semantic space. The framework incorporates condition-aware and task-aware calibration, along with co-teacher distillation tailored for healthcare applications. These advancements significantly enhance the distinguishability and task awareness of encoded representations, as well as the code-level alignment between textual and CO signals. Extensive experiments validate the efficacy of our approach. However, our model has limitations, including the need to integrate modalities beyond text, which will be explored in future work.

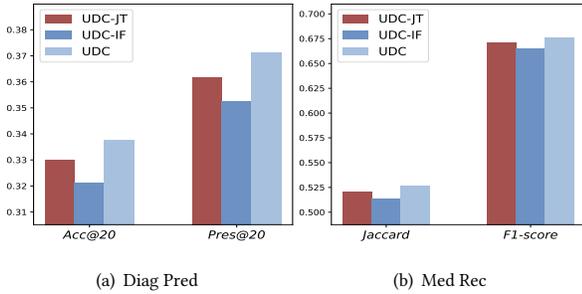
References

- [1] Zafar Ali, Yi Huang, Irfan Ullah, Junlan Feng, et al. 2023. Deep learning for medication recommendation: a systematic survey. *Data Intelligence* 5, 2 (2023), 303–354.
- [2] Kamal Berahmand, Fatemeh Daneshfar, Elaheh Sadat Salehi, Yuefeng Li, and Yue Xu. 2024. Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review* 57, 2 (2024), 28.
- [3] Suman Bhoi, Mong Li Lee, Wynne Hsu, Hao Sen Andrew Fang, and Ngiap Chuan Tan. 2021. Personalizing medication recommendation with a graph-based approach. *ACM Transactions on Information Systems (TOIS)* 40, 3 (2021), 1–23.
- [4] Tanujit Chakraborty, Ujjwal Reddy KS, Shradha M Naik, Madhurima Panja, and Bayapureddy Manvitha. 2024. Ten years of generative adversarial nets (GANs): a survey of the state-of-the-art. *Machine Learning: Science and Technology* 5, 1 (2024), 011001.
- [5] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 787–795.
- [6] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* 29 (2016).
- [7] Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G Lee, Vikram Deshpande, Joseph Schwab, Michael H Lev, Ramon G Gonzalez, Michael S Gee, and Synho Do. 2023. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering* 7, 6 (2023), 711–718.
- [8] MICHAEL H Criqui, A Fronck, MRI Klauer, E Barrett-Connor, and S Gabriel. 1985. The sensitivity, specificity, and predictive value of traditional clinical evaluation of peripheral arterial disease: results from noninvasive testing in a defined population. *Circulation* 71, 3 (1985), 516–522.
- [9] Anne AH de Hond, Artuur M Leeuwenberg, Lotty Hooft, Ilse MJ Kant, Steven WJ Nijman, Hendrikus JA van Os, et al. 2022. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ digital medicine* 5, 1 (2022), 2.
- [10] Linda Ekerljung, Apostolos Bossios, Jan Lötvall, Anna-Carin Olin, Eva Rönmark, Göran Wennergren, Kjell Torén, and Bo Lundbäck. 2011. Multi-symptom asthma as an indication of disease severity in epidemiology. *European Respiratory Journal* 38, 4 (2011), 825–832.
- [11] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020*. 530–540.
- [12] Miguel Garcia-Argibay, Lin Li, Ebba Du Rietz, Le Zhang, Honghui Yao, Johan Jendle, Josep A Ramos-Quiroga, Marta Ribases, Zheng Chang, Isabell Brikell, et al. 2023. The association between type 2 diabetes and attention-deficit/hyperactivity disorder: A systematic review, meta-analysis, and population-based sibling study. *Neuroscience & biobehavioral reviews* 147 (2023), 105076.
- [13] Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. 2021. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications* 12, 1 (2021), 711.
- [14] Pengcheng Jiang, Cao Xiao, Adam Richard Cross, and Jimeng Sun. 2024. GraphCare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs. In *The Twelfth International Conference on Learning Representations*.
- [15] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. 2023. Diffusionret: Generative text-video retrieval with diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2470–2481.
- [16] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. MIMIC-iv. *PhysioNet*. (2020).
- [17] Alistair EW Johnson et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.
- [18] Kibum Kim, Dongmin Hyun, Sukwon Yun, and Chanyoung Park. 2023. MELT: Mutual Enhancement of Long-Tailed User and Item for Sequential Recommendation. In *Proceedings of the 46th international ACM SIGIR conference on Research and development in information retrieval*. 68–77.
- [19] Taeri Kim, Jiho Heo, Hongil Kim, et al. 2024. VITA: Carefully Chosen and Weighted Less' Is Better in Medication Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8600–8607.
- [20] Tzu-Lin Kuo, Tzu-Wei Chiu, Tzung-Sheng Lin, Sheng-Yang Wu, Chao-Wei Huang, and Yun-Nung Chen. 2024. A Survey of Generative Information Retrieval. *arXiv e-prints* (2024), arXiv:2406.
- [21] Doyup Lee, Chiheon Kim, et al. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11523–11532.
- [22] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2024. From matching to generation: A survey on generative information retrieval. *arXiv preprint arXiv:2404.14851* (2024).
- [23] Xiang Li, Shunpan Liang, Yulei Hou, and Tengfei Ma. 2024. StratMed: Relevance stratification for biomedical entities for sparsity on medication recommendation. *Knowledge-Based Systems* 284 (2024), 111239.
- [24] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023. Generative retrieval for conversational question answering. *Information Processing & Management* 60, 5 (2023), 103475.
- [25] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4228–4238.
- [26] Sicen Liu, Xiaolong Wang, Jingcheng Du, Yongshuai Hou, Xianbing Zhao, Hui Xu, Hui Wang, Yang Xiang, and Buzhou Tang. 2023. SHAPE: A Sample-adaptive Hierarchical Prediction Network for Medication Recommendation. *IEEE Journal of Biomedical and Health Informatics* (2023).
- [27] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering* 35, 1 (2021), 857–876.
- [28] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 647–656.
- [29] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics* 23, 6 (2022), bbac409.
- [30] Hang Lv, Zehai Chen, Yacong Yang, Guofang Ma, Tan Yanchao, and Carl Yang. 2024. BoxCare: A Box Embedding Model for Disease Representation and Diagnosis Prediction in Healthcare Data. In *Companion Proceedings of the ACM on Web Conference 2024*. 1130–1133.
- [31] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1903–1911.
- [32] Fenglong Ma, Yaqing Wang, Jing Gao, et al. 2020. Rare Disease Prediction by Generating Quality-Assured Electronic Health Records. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 514–522.
- [33] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2016. Deepr: a convolutional net for medical records. *IEEE journal of biomedical and health informatics* 21, 1 (2016), 22–30.
- [34] Tom J Pollard, Alistair EW Johnson, et al. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data* 5, 1 (2018), 1–13.
- [35] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2024. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems* 36 (2024).
- [36] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* 32 (2019).
- [37] Wolfgang Sadec, Danxin Wang, Katherine Hartmann, and Amanda Ewart Toland. 2023. Pharmacogenomics: driving personalized medicine. *Pharmacological reviews* 75, 4 (2023), 789–814.
- [38] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1126–1133.
- [39] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2024. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems* 36 (2024).
- [40] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
- [41] Ioannis S Vizirianakis. 2011. Nanomedicine and personalized medicine toward the application of pharmacotyping in clinical practice to improve drug-delivery outcomes. *Nanomedicine: Nanotechnology, Biology and Medicine* 7, 1 (2011), 11–17.
- [42] Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023. Optimized glyemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine* 29, 10 (2023), 2633–2642.
- [43] Liping Wang, Qiang Liu, Mengqi Zhang, Yaxuan Hu, Shu Wu, and Liang Wang. 2023. Stage-aware hierarchical attentive relational network for diagnosis prediction. *IEEE Transactions on Knowledge and Data Engineering* (2023).
- [44] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2022. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2022), 4425–4445.
- [45] Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional generation net for medication recommendation. In *Proceedings of the ACM Web*

1045		Conference 2022. 935–945.	1103
1046	[46]	Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. 2024. Achieving cross modal generalization with multimodal unified representation. <i>Advances in Neural Information Processing Systems</i> 36 (2024).	1104
1047			1105
1048	[47]	Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tiejian Liu. 2023. A survey on non-autoregressive generation for neural machine translation and beyond. <i>IEEE Transactions on Pattern Analysis and Machine Intelligence</i> 45, 10 (2023), 11407–11427.	1106
1049			1107
1050			1108
1051	[48]	Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In <i>Proceedings of the ACM Web Conference 2023</i> . 2819–2830.	1109
1052			1110
1053	[49]	Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P Danek, and Jimeng Sun. 2023. Pyhealth: A deep learning toolkit for healthcare applications. In <i>Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining</i> . 5788–5789.	1111
1054			1112
1055			1113
1056	[50]	Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2021. Change matters: Medication change prediction with recurrent residual networks. <i>arXiv preprint arXiv:2105.01876</i> (2021).	1114
1057			1115
1058	[51]	Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. Safe-Drug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations. In <i>30th International Joint Conference on Artificial Intelligence, IJCAI 2021</i> . International Joint Conferences on Artificial Intelligence, 3735–3741.	1116
1059			1117
1060			1118
1061	[52]	Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. 2023. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In <i>Proceedings of the ACM Web Conference 2023</i> . 4075–4085.	1119
1062			1120
1063			1121
1064	[53]	Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medpath: Augmenting health risk prediction via medical knowledge paths. In <i>Proceedings of the Web Conference 2021</i> . 1397–1409.	1122
1065			1123
1066	[54]	Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, and Yixin Chen. 2019. D-vae: A variational autoencoder for directed acyclic graphs. <i>Advances in neural information processing systems</i> 32 (2019).	1124
1067			1125
1068			1126
1069	[55]	Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2023. Deep long-tailed learning: A survey. <i>IEEE Transactions on Pattern Analysis and Machine Intelligence</i> 45, 9 (2023), 10795–10816.	1127
1070			1128
1071	[56]	Chuang Zhao, Hongke Zhao, Xiaofang Zhou, and Xiaomeng Li. 2024. Enhancing Precision Drug Recommendations via In-depth Exploration of Motif Relationships. <i>IEEE Transactions on Knowledge and Data Engineering</i> (2024).	1129
1072			1130
1073	[57]	Xiangyu Zhao, Maolin Wang, Xinjian Zhao, Jiansheng Li, Shucheng Zhou, Dawei Yin, Qing Li, Jiliang Tang, and Ruocheng Guo. 2023. Embedding in recommender systems: A survey. <i>arXiv preprint arXiv:2310.18608</i> (2023).	1131
1074			1132
1075	[58]	Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, and Xiangnan He. 2024. Leave No Patient Behind: Enhancing Medication Recommendation for Rare Disease Patients. <i>47th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> (2024).	1133
1076			1134
1077	[59]	Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. 2022. Movq: Modulating quantized vectors for high-fidelity image generation. <i>Advances in Neural Information Processing Systems</i> 35 (2022), 23412–23425.	1135
1078			1136
1079			1137
1080			1138
1081			1139
1082			1140
1083			1141
1084			1142
1085			1143
1086			1144
1087			1145
1088			1146
1089			1147
1090			1148
1091			1149
1092			1150
1093			1151
1094			1152
1095			1153
1096			1154
1097			1155
1098			1156
1099			1157
1100			1158
1101			1159
1102			1160

Table 4: Mathematical Notations.

Notations	Descriptions
\mathcal{U}	any ehr dataset
\mathcal{N}	patient set
$\mathcal{D}, \mathcal{P}, \mathcal{M}$	diag. proc, and medication set
$\mathbf{d}, \mathbf{p}, \mathbf{m}$	multi-hot code of diag, proc, and medication
t	t -th visit
\mathcal{T}	length of visit
T	text description
$\mathcal{F}_{co}, \mathcal{F}_{te}$	pcm model, plm model
$\mathbf{E}, \tilde{\mathbf{E}}$	collaborative embedding, text embedding
C_l	l -th level codebook
c_i, \mathbf{c}_i	i code index, i -th code
ϕ_{co}, ϕ_{te}	mlp encoder
$\mathbf{z}, \tilde{\mathbf{z}}$	disease representation (CO, text)
\mathbf{f}	condition representation
$\varphi_\gamma, \varphi_\beta$	transformation matrix
\mathbf{s}	next-visit target representation
\mathbf{o}	holistic view of disease representation
\mathbf{b}	relation representation from another view
κ	decay rate
\mathbf{n}	norm vector
sg	stop gradient
α	commitment weight
ψ_{co}, ψ_{te}	mlp decoder
\mathcal{L}	loss function

**Figure 8: Different training methods (MIMIC-III). Please note that UDC refers to the strategy utilized in the manuscript.**

A Mathematical Notations

Annotations for essential symbols are shown in Table 4.

B Diverse Training Methods

We also experiment with various training methods, such as joint training (θ, Θ) and inference without fine-tuning, as depicted in Figure 8. Formally, UDC-JT trains PCM and DRL simultaneously, and we observe that this model initially focuses on learning collaborative signals, leading to DRL training collapse. In contrast, UDC-IF skips fine-tuning and directly performs inference. However, since $\mathcal{F}_{co}(\cdot)$ does not fully capture the interaction patterns between rare and common diseases, improvements stem primarily from the integration of textual semantic information. From UDC, it

is evident that learning these interaction patterns plays a critical role in enhancing the model’s overall performance.

C Algorithm

The algorithm flow is shown in Algorithm 1.

Algorithm 1 The Algorithm of UDC

Input: EHR \mathcal{U} , Textual Knowledge $T(\cdot)$, Rare threshold η ;
Output: PCM parameters θ , DRL parameter Θ ;

- 1: **Stage 1: Backbone Training** ▷ Tuning θ
- 2: PCM training $\mathbf{e}_d \in E_D, \mathbf{e}_p \in E_P, \mathbf{e}_m \in E_M$;
- 3: PLM Initialization $\tilde{\mathbf{e}}_d \in \tilde{E}_D, \tilde{\mathbf{e}}_p \in \tilde{E}_P, \tilde{\mathbf{e}}_m \in \tilde{E}_M$;
- 4: **Stage 2: DRL Training** ▷ Frozen \mathbf{E} & $\tilde{\mathbf{E}}$, Tuning Θ
- 5: Split disease into $\mathcal{D}_{com}, \mathcal{D}_{rar}$ using η ;
- 6: **while** not converged **do**
- 7: Sample disease d from \mathcal{D}_{com} ;
- 8: Extract PCM & PLM embedding $\mathbf{e}_d, \tilde{\mathbf{e}}_d$ in Eq. 1-3;
- 9: Obtain discrete representation \mathbf{z}_d and $\tilde{\mathbf{z}}_d$ in Eq. 5;
- 10: Condition-aware calibration in Eq. 7;
- 11: Task-aware calibration in Eq. 10;
- 12: Co-teacher distillation for codebook in Eq. 12-13;
- 13: Optimization in Eq. 14;
- 14: Update the parameters;
- 15: **end while**
- 16: **Stage 3: Fine-tuning** ▷ Frozen Θ , Tuning θ
- 17: Obtain enhanced disease representation $\hat{\mathbf{e}}_d$ in Eq. 15;
- 18: Fine-tuning θ using Eq. 2;
- 19: **return** Parameters θ & Θ ;

D Dataset Statistics

MIMIC-III is a widely utilized dataset containing EHRs from over 40,000 patients in critical care. MIMIC-IV, the successor to MIMIC-III, expands on this with data from over 70,000 admissions, reflecting more recent practices and broader patient demographics. eICU comprises health data from over 200,000 patients across various ICU settings in the United States, offering extensive coverage of diverse clinical environments and treatment modalities. We present the dataset statistics after pre-processing [49, 56] in Table 5.

E Metric Definitions

The evaluation metrics outlined in the paper are defined as follows,

$$\text{Jaccard} = \frac{|y_{\text{pred}} \cap y_{\text{true}}|}{|y_{\text{pred}} \cup y_{\text{true}}|}, \quad (17)$$

where y_{pred} is the predicted set and y_{true} is the true set.

$$\text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (18)$$

where precision and recall are defined as: precision = $\frac{\text{tp}}{\text{tp} + \text{fp}}$ and recall = $\frac{\text{tp}}{\text{tp} + \text{fn}}$, where tp is the number of true positives, fp is the number of false positives, and fn is the number of false negatives.

$$\text{AUROC} = \int_0^1 \text{tpr}(\text{fpr}) d(\text{fpr}), \quad (19)$$

where tpr (True Positive Rate) and fpr (False Positive Rate) are calculated at different thresholds.

Table 5: Data Statistics across all datasets (Diag Pred || Med Rec). Due to task-specific preprocessing variations, we present data statistics for all tasks. # means the number of.

Items	MIMIC-III	MIMIC-IV	eICU	MIMIC-III	MIMIC-IV	eICU
# of patients / # of visits	6,164 / 9,693	26,697 / 99,668	8,853 / 10,188	35,707 / 44,399	46,187 / 154,962	114,473 / 124,564
diag. / prod. / med. set size	4,017 / 1,274 / 192	16,906 / 9,026 / 199	1,326 / 422 / 1,411	6,662 / 1,978 / 197	19,438 / 10,790 / 200	1,670 / 461 / 1,411
avg. # of visits	1.5725	3.7333	1.1508	1.2434	3.3551	1.0882
avg. # of diag per visit	27.7807	58.2390	10.1569	17.7373	48.9516	7.6574
avg. # of prod per visit	7.7473	9.7644	32.6515	6.1718	8.7626	27.9025
avg. # of drug per visit	29.6780	24.6252	15.7981	27.1113	23.8334	17.2664

$$\text{AUPRC} = \int_0^1 \text{precision}(\text{recall}) d(\text{recall}), \quad (20)$$

where precision and recall are evaluated at different thresholds.

$$\text{Pres@K} = \frac{\sum_{i=1}^K \text{relevant}_i}{K}, \quad (21)$$

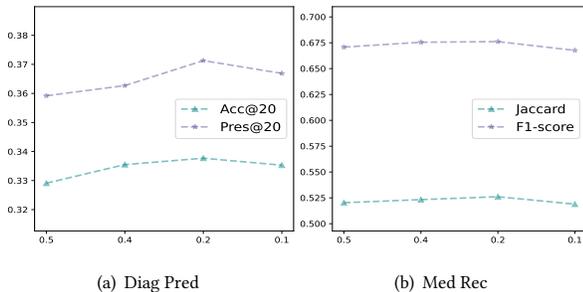
where relevant_i is 1 if the i -th predicted item is relevant (i.e., a true positive), and 0 otherwise.

$$\text{Acc@K} = \frac{\sum_{i=1}^k \mathbf{1}(y_{\text{pred},i} = y_{\text{true},i})}{K}, \quad (22)$$

where $\mathbf{1}$ is the indicator function, equal to 1 if $y_{\text{pred},i} = y_{\text{true},i}$, and 0 otherwise. Pres@K measures performance at the visit level, while Acc@K evaluates performance at the code level [43]. For all metrics, higher values are preferable.

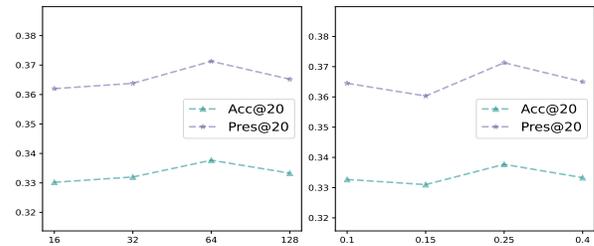
F Further Analysis

We further discuss several key hyperparameters.

**Figure 9: Performance under different ratios. (MIMIC-III)**

Rare Ratio η . Our results, as shown in Figure 9, indicate that UDC achieves the best performance when $\eta = 20\%$. When η is too low, the DRL may not be well-trained from a limited CO-Text pair, making it difficult to obtain semantic alignment between CO and textual spaces. Conversely, with a very high value for η , UDC does not yield significant improvement. The restricted absolute quantity results in fewer rare disease entity adjustments, exerting minimal influence on the comprehensive sequence representation.

Codebook Size $|C_l|$. The codebook size is a critical hyperparameter in the VQ-VAE architecture [21]. A larger codebook size allows the VQ-VAE to capture a richer set of discrete latent features, enabling more detailed and expressive reconstructions of the disease symptoms. However, this comes at the cost of increased computational complexity and potential overfitting, especially when working with limited training data. In contrast, a smaller codebook size can lead to more robust and generalized representations but may struggle

**Figure 10: (a) Performance under different codebook sizes. (b) Performance under different commitment weights. We show the results on MIMIC-III (Diag Pred).**

to represent the full complexity of the input distribution. Experimentally, we set $|C_l| = 64$.

Commitment Weight α . α is a crucial parameter influencing the quality of absolute code representation and alignment. A larger value enhances the similarity between the encoded representation and the discrete representation, thereby improving cross-domain alignment. However, increasing α may reduce the emphasis on the reconstruction target, potentially leading to negative effects. Experimentally, we set $\alpha = 0.25$.

G Case Study: Real Prediction

To intuitively demonstrate the superiority of UDC, we present the medication recommendations for a randomly selected patient. Specifically, UDC achieves a significantly higher Jaccard compared to the other baselines. This indicates that UDC can generate diagnostic and treatment suggestions that are much closer to the clinically validated outcomes and better distinguish between positive and negative samples. Furthermore, F1-score generated by our model is also higher compared to RAREMed. This finding suggests that instead of relying on broad recommendations to enhance performance metrics, our framework offers improved recommendations that effectively balance sensitivity and specificity [8].

Table 6: Example recommendation result.

Method	Recommended Med Set
Ground-Truth	TP: ['B05X', 'B01A', 'A12B', 'C07A', 'A06A', 'C10A', 'N02B', 'A03B', 'C09A', 'N06A', 'A04A', 'C09C']
RAREMed	TP: ['A03B', 'A06A', 'A12B', 'B01A', 'B05X', 'C07A', 'C09A', 'C10A', 'N02B']
F1-score:0.7500	FN: ['A04A', 'C09C', 'N06A']
Jaccard:0.6000	FP: ['A02B', 'A12C', 'N02A']
UDC	TP: ['A03B', 'A06A', 'A12B', 'B01A', 'B05X', 'C07A', 'C09A', 'C10A', 'N02B']
F1-score:0.8181	FN: ['A04A', 'C09C', 'N06A']
Jaccard:0.6923	FP: ['A12C']