# Multi-stage Retrieve and Re-rank Model for Automatic Medical Coding Recommendation

**Xindi Wang**[1,2], **Robert E. Mercer**[1], **Frank Rudzicz**[2,3,4]
[1] Department of Computer Science, University of Western Ontario, Canada
[2] Vector Institute for Artificial Intelligence, Canada
[3] Faculty of Computer Science, Dalhousie University, Canada
[4] Department of Computer Science, University of Toronto, Canada
xwang842@uwo.ca, mercer@csd.uwo.ca, frank@dal.ca

## Abstract

The International Classification of Diseases (ICD) serves as a definitive medical classification system encompassing a wide range of diseases and conditions. The primary objective of ICD indexing is to allocate a subset of ICD codes to a medical record, which facilitates standardized documentation and management of various health conditions. Most existing approaches have suffered from selecting the proper label subsets from an extremely large ICD collection with a heavy long-tailed label distribution. In this paper, we leverage a multi-stage "retrieve and re-rank" framework as a novel solution to ICD indexing, via a hybrid discrete retrieval method, and re-rank retrieved candidates with contrastive learning that allows the model to make more accurate predictions from a simplified label space. The retrieval model is a hybrid of auxiliary knowledge of the electronic health records (EHR) and a discrete retrieval method (BM25), which efficiently collects high-quality candidates. In the last stage, we propose a label co-occurrence guided contrastive re-ranking model, which re-ranks the candidate labels by pulling together the clinical notes with positive ICD codes. Experimental results show the proposed method achieves state-of-the-art performance on a number of measures on the MIMIC-III benchmark.

## 1 Introduction

Electronic health records[1] (EHRs) contain a comprehensive repository of essential administrative and clinical data pertinent to a person's care within a specific healthcare provider setting. In order to conduct meaningful statistical analysis, these EHR data are annotated with structured codes in a classification system known as *medical codes*. The International Classification of Diseases[2] (ICD) is one of the most widely-used coding systems, and it provides a taxonomy of classes, each uniquely identified by a code assigned to an episode of patient care.

The task of medical coding associates ICD codes with EHR documents. The *status quo* of assigning medical codes is a manual process, which is labour-intensive, time-consuming, and error-prone (Xie and Xing, 2018). To reduce coding errors and cost, the demand for automated medical coding has become imperative. Previous deep learning approaches regarded medical coding as an extreme multi-label text classification problem (Shi et al., 2017; Mullenbach et al., 2018; Baumel et al., 2018; Xie et al., 2019; Yuan et al., 2022), where an encoder is typically employed to learn the representations of the clinical notes and a label-specific binary classifier is subsequently constructed on top of the encoder for label predictions. However, some remaining difficulties have still posed immense challenges. First, clinical documents are lengthy (containing on average 1596 words in the MIMIC-III dataset) and noisy (including terse abbreviations, symbols, and misspellings). Second, the label set is extremely large and complex; for instance, in the $10^{th}$ ICD edition, there are over 130,000 codes[3]. Third, the distribution of ICD codes is extremely long-tailed; while some ICD codes occur frequently, many others seldom appear, if at all, because of the rarity of the diseases. For instance, among the 942 unique 3-digit ICD codes in the MIMIC-III dataset (Johnson et al., 2016), the ten most common codes account for 26% of all code occurrences and the 437 least common codes account for only 1% of occurrences (Bai and Vucetic, 2019). To address the aforementioned challenges, we propose a novel multi-stage retrieve and re-rank framework, where the goal is to first generate a curated ICD list and then

[1] https://www.cms.gov/Medicare/E-Health/EHealthRecords

[2] https://www.who.int/standards/classifications/classification-of-diseases

[3] https://www.cdc.gov/nchs/icd/icd10cm_pcs.htm



*Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4881–4891


Figure 1: An example of a medical record from the MIMIC-III dataset which includes the discharge summary, assigned ICD codes and auxiliary knowledge. We colour each code and its corresponding mentions in the discharge summary and auxiliary knowledge. We use the auxiliary knowledge of the notes to retrieve the candidate subset of the label space.

provide suggested ICD codes for a given medical record. In contrast to prior approaches, for instance, CAML (Mullenbach et al., 2018), MultiResCNN (Li and Yu, 2020) and KEPTLongformer (Yang et al., 2022), that primarily consider ICD indexing as a multi-label text classification task, we introduce a new perspective that conceptualizes the task as a recommendation problem. More precisely, we first conduct a two-stage retrieval process leveraging auxiliary knowledge and BM25 to obtain a small subset of candidate ICD codes from the large number of labels to alleviate issues caused by the label set and imbalanced label distribution. EHR auxiliary knowledge holds significant potential, but it has often been underutilized in prior studies. In addition to clinical texts, our focus centers on two code terminologies: Diagnosis-Related Group codes[4] (DRG) and Current Procedural Terminology codes[5] (CPT), as well as patient prescribed medications. These external sources can serve as robust indicators for predicting ICD codes. For instance, within a drug prescription, the presence of a medication like "Namenda" can strongly imply a likelihood of Alzheimer's disease, as depicted in Figure 1. Subsequently, we design a re-ranking model via co-occurrence guided contrastive learning to refine the candidate set, which can deal with lengthy clinical notes and generate semantically meaningful representations via the pre-trained language model and leverage code co-occurrence to generate co-occurrence-aware label representations. The co-occurrence of codes in clinical texts yields valuable insights into the interconnections among different diseases or conditions. As illustrated in Figure 1, the code for "Dementia in conditions classified elsewhere without behavioral disturbance" (294.10) can be easily found in the text; however, inferring the code "Alzheimer's disease" (331.0) presents a more intricate challenge with less explicit clues. Fortunately, a robust association exists between these two diseases, with "Alzheimer's disease" serving as a prevalent cause of "dementia". This linkage can be effectively captured as these two diseases frequently co-occur within the clinical notes. This empowers us to gain a deeper understanding of the contexts, which could mitigate the limitation of long-tailed label distributions as rare labels might be suggested based on these relationships. We train the re-ranking model via contrastive learning as it has strong discriminative power that can extract features uniquely associated with each class, which empowers the model to make more accurate recommendations.

To summarize, the major contributions of this paper are:

- We formalize the medical coding task as a recommendation problem and present a novel multi-stage retrieve and re-rank framework to make more accurate predictions by ruling out the irrelevant codes before ranking, rather than making direct predictions on the entire large label set.
- To address the large label set and long-tailed distribution issues, in the two-stage retrieval process we use external knowledge and BM25 to retrieve a subset of candidate labels from the large label space. We further leverage the code co-occurrence in the re-ranking stage to capture the internal connections among the codes.
- We apply contrastive learning in the re-ranking stage. It effectively pulls together the

[4] https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/MS-DRG-Classifications-and-Software

[5] https://www.ama-assn.org/amaone/cpt-current-procedural-terminology

representations of a clinical note and its corresponding golden truth labels, which allows the model to make more accurate predictions.

## 2 Related Work

The automatic ICD indexing task is well established in the healthcare domain. Extensive research using deep learning has been dedicated to ICD indexing, including recurrent-based neural networks (RNNs), convolution-based neural networks (CNNs), and their variations (Mullenbach et al., 2018; Li and Yu, 2020; Shi et al., 2017; Xie and Xing, 2018). These architectures are able to extract and categorize semantic features, reducing the need for medical domain expertise during the traditional feature selection stage seen in conventional algorithms (Teng et al., 2023). The ICD indexing task is formulated as a multi-label classification problem in these approaches. Mullenbach et al. (2018) introduced a combination of CNN with an attention mechanism to effectively capture pertinent information within clinical texts for each ICD code. Building on this foundation, Xie et al. (2019) enhanced the CNN attention model by integrating a multi-scale feature attention technique. Many CNN variants were subsequently introduced to address the challenges posed by lengthy and noisy clinical texts, including MultiResCNN (Li and Yu, 2020), DCAN (Ji et al., 2020), and EffectiveCAN (Liu et al., 2021). RNN-based models, renowned for their capacity to capture contextual information across input texts, have also been widely used for ICD indexing. Shi et al. (2017) proposed a character-aware Long Short-Term Memory (LSTM) recurrent network to learn the underlying representations of clinical texts. Xie and Xing (2018) introduced a tree-of-sequences LSTM architecture alongside adversarial learning to capture hierarchical relationships among ICD codes. Additionally, Baumel et al. (2018) presented a Hierarchical Attention-Bidirectional Gated Recurrent Unit (HA-GRU) model, facilitating document labeling by identifying sentences relevant to each ICD code. LAAT (Vu et al., 2020) used a bidirectional Long-Short Term Memory (BiLSTM) encoder and a customized label-wise attention mechanism to cultivate label-specific vectors across distinct clinical text fragments.

To address the hierarchical relationships intrinsic to ICD codes, Graph Convolutional Neural Networks (GCNNs) (Kipf and Welling, 2017) have emerged as a powerful tool. Rios and Kavuluru (2018) and Xie et al. (2019) used GCNNs to capture both the hierarchical interplay among ICD codes and the semantic information specific to each code. HyperCore (Cao et al., 2020) took a comprehensive approach by considering both code hierarchy and code co-occurrence, employing GCNNs to learn code representations within the co-graph.

Incorporating external knowledge beyond ICD code information has also gained traction. Bai and Vucetic (2019) introduced a Knowledge Source Integration (KSI) model that integrates external knowledge from Wikipedia. This integration calculated matching scores between clinical notes and disease-related Wikipedia documents, in order to enrich the available information for ICD predictions. Additionally, Yuan et al. (2022) proposed a Multiple Synonym Matching Network (MSMN) to use synonyms of ICD codes, enhancing the quality of code representation learning. Expanding on this, Yang et al. (2022) integrated a pre-trained language model with three domain-specific knowledge sources: code hierarchy, synonyms, and abbreviations. This fusion of knowledge sources contributes significantly to the performance of ICD classification.

## 3 Method

### 3.1 A Multi-stage Framework

We formulate the medical coding task as a recommendation task given medical records $\mathcal{D} = \{d_1, d_2, ..., d_N\}$ and a set of ICD codes $\mathcal{Y} = \{y_1, y_2, ..., y_L\}$ with associated external auxiliary knowledge $\mathcal{K}$. We construct the label information as a graph structure $\mathcal{G}$, using code co-occurrence relations, and we train a multi-stage recommender system $\mathcal{R}$, based on the text information $\mathcal{D}$, constructed label information $\mathcal{G}$, and the external auxiliary knowledge $\mathcal{K}$. The system $\mathcal{R}$ needs to predict the relevant labels given a document $d \notin \mathcal{D}$.

In this section, we present a multi-stage retrieve and re-rank framework for ICD indexing, which is shown in Figure 2. Our model is composed of a two-stage retrieval process that uses auxiliary knowledge of the EHR and BM25 to obtain a shortened candidate list, and a re-ranking process that conducts code co-occurrence guided contrastive learning to further improve the recommended ICD list.

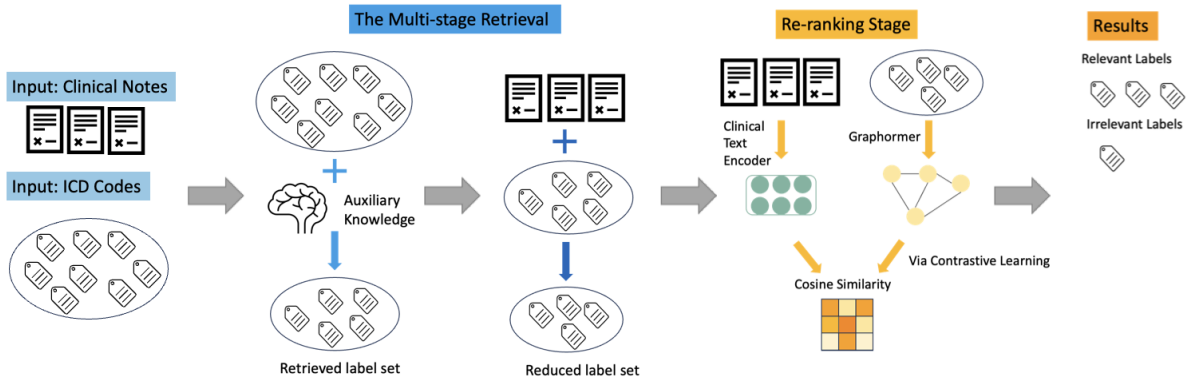


Figure 2: Overview of the proposed multi-stage retrieve and re-rank framework. The model first leverages auxiliary knowledge and BM25 to retrieve a candidate list from the full label space, then uses a re-rank model that leverages the code co-occurrence guided contrastive learning to generate the final relevant labels.

### 3.2 The Retrieval Stage

**Using Auxiliary Knowledge** To retrieve the candidate list using auxiliary knowledge, we incorporate insights from three external sources of knowledge: diagnosis-related group (DRG) codes, current procedural terminology (CPT) codes, and medications prescribed to patients. DRG codes are used by hospitals and healthcare providers to classify patients into groups based on their diagnosis, treatment, and length of stay. These codes are used for reimbursement purposes, and they help determine the amount that healthcare providers are remunerated for their services. DRG codes are further classified into medical DRGs (which exclude operating room procedures) and surgical DRGs. CPT codes are used to describe medical procedures and services provided by healthcare providers. They provide a standardized way of documenting and billing for medical services. CPT codes are used by insurance companies to determine reimbursement rates for healthcare providers. Such code terminologies significantly contribute to the refinement of ICD indexing. Moreover, the medications prescribed to patients offer a wealth of predictive information for ICD codes. These prescriptions often mark the conclusion of a patient's care episode. As patients approach the conclusion of their treatment, the prescribed medications serve a critical role in managing their conditions. Consequently, these medications emerge as potent indicators of underlying health conditions or diagnoses. Their inclusion in the retrieval process greatly enhances the accuracy and relevance of the corresponding ICD code recommendations. The aforementioned auxiliary knowledge, such as DRG codes, CPT codes, and drug prescriptions, typically appears in the EHR data and is readily accessible.

Given a clinical note $d$, we retrieve the candidate ICD list by calculating the auxiliary knowledge and label co-occurrence matrix using conditional probabilities, i.e., $P(y_i \mid k_j)$, which denote the probabilities of occurrence of ICD $y_i$ when auxiliary knowledge $k_j$ appears.

$$P(y_i \mid k_j) = \frac{C_{y_i \cap k_j}}{C_{k_j}}, \quad (1)$$

where $C_{y_i \cap k_j}$ denotes the number of co-occurrences of $y_i$ and $k_j$, and $C_{k_j}$ is the number of occurrences of $k_j$ in the training set. To avoid the noise of rare co-occurrences, a threshold $\eta$ filters noisy correlations. $\tilde{K}_j$ denotes the selected ICD set for auxiliary knowledge $j$.

$$\tilde{K}_j = \{y_i | P(y_i|k_j) > \eta,\ i = 1, ..., L\}, \quad (2)$$

where $L$ is the total number of ICD codes in the label set, and $\eta = 0.005$. We then join the ICD codes retrieved from the auxiliary knowledge co-occurrences for the DRG codes, CPT codes and prescribed drugs to form the candidate ICD subset $\mathcal{C}_{\text{auxiliary}}$:

$$\mathcal{C}_{\text{auxiliary}}(d) = \tilde{K}_{\text{DRG}}(d) \cup \tilde{K}_{\text{CPT}}(d) \cup \tilde{K}_{\text{drug}}(d), \quad (3)$$

where $\mathcal{C}_{\text{auxiliary}} \subseteq \mathcal{Y}$.

**Using BM25** The retrieval stage using auxiliary knowledge incorporates the co-relations between ICD codes and external knowledge, but ignores

the relationship between clinical texts and labels. To increase the recall of the retrieval stage, we adopt BM25 (Robertson and Walker, 1994) to allow lexical matching between the medical documents and labels on the retrieved candidate list $\mathcal{C}_{\text{auxiliary}}$. Given a medical record $d$ and an ICD code $y$, the score between $d$ and $y$ is calculated as:

$$\text{BM25}(d, y) = \sum_{w \in d \cap t_y} \text{IDF}(w) \frac{\text{TF}(w,t_y)\cdot(k_1+1)}{\text{TF}(w,t_y)\cdot k_1(1-b+b\frac{|\mathcal{Y}|}{avgdl})}, \quad (4)$$

and

$$avgdl = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} |t_y|, \quad (5)$$

where $t_y$ represents the words in the label descriptors, $|\mathcal{Y}|$ is the length of the label descriptors in words, *avgdl* is the average length of text information in the label.

When the BM25 score between $d$ and $y_i$ exceeds a certain threshold $\theta$, we add $y_i$ as a candidate of $d$:

$$\mathcal{C}_{\text{BM25}}(d) = \{y_i | \text{BM25}(d, y_i) > \theta, y_i \in \mathcal{C}_{\text{auxiliary}}\}, \quad (6)$$

where $\theta = 200$. Given a clinical note $d$, its candidate ICD set is first generated by using the auxiliary knowledge in the retrieval stage and then reduced by using BM25, where $\mathcal{C}_{\text{BM25}} \subseteq \mathcal{C}_{\text{auxiliary}}$ and $\mathcal{C}_{\text{auxiliary}} \subseteq \mathcal{Y}$.

### 3.3 The Re-ranking Stage

**Clinical Text Encoder** Encouraged by the success of the pre-trained language model Longformer (Beltagy et al., 2020) in dealing with longer texts, we use Clinical-Longformer (Li et al., 2023), specifically pre-trained in the medical domain, as a text encoder. Given a medical document $d$ as input that consists of a sequence of tokens:

$$d = \{[\texttt{CLS}], x_1, x_2, ..., x_{n-2}, [\texttt{SEP}]\}, \quad (7)$$

where [CLS] and [SEP] are two special tokens that indicate the beginning and end of the sequence, and $n$ is the sequence length, the Clinical-Longformer encodes the tokens and outputs the hidden representations for each token:

$$H_{\text{hidden}} = \text{ClinicalLongformer}(d), \quad (8)$$

where $H_{\text{hidden}} \in \mathbb{R}^{n \times h_e}$, and $h_e$ is the hidden size. Following previous work (Wang et al., 2022; Yang et al., 2022), we use the hidden state of the [CLS] token to represent the document, which is the first token of $H_{\text{hidden}}$, denoted as $H_{\text{T}}$.

**Label Encoder** The occurrence of two ICD codes together in clinical texts frequently indicates a simultaneous presence or a causal connection between specific diseases. This implies that the codes representing these interconnected diseases often manifest together within clinical notes. We employ a Graphormer (Ying et al., 2021) to incorporate the co-occurrence relationships among ICD codes. Unlike the original GNN, Graphormer models graphs using Transformer layers (Vaswani et al., 2017) with spatial encoding and edge encoding, which could effectively encode the structural information (i.e., code co-occurrence) of a graph into the model. We create a directed code co-occurrence graph $\mathcal{G} = (\mathcal{Y}, \mathcal{E})$, where node set $\mathcal{Y}$ is the labels and edge set $\mathcal{E}$ denotes the co-occurrence relations. This graph is constructed using the code co-occurrence matrix, which has been used as the edge matrix for the graph. We create the code co-occurrence matrix by using the correlated relationship between labels based on conditional probabilities. This approach encapsulates the interdependence between various ICD codes in a quantifiable manner, offering valuable insights into the underlying connections among disease codes within the clinical texts. To be more specific, we calculate the probability of occurrence of label $y_j$ when label $y_i$ appears as follows:

$$P(y_j \,|\, y_i) = \frac{C_{y_i \cap y_j}}{C_{y_i}} \quad (9)$$

where $C_{y_j \cap y_i}$ denotes the number of co-occurrences of $y_i$ and $y_j$, and $C_{y_i}$ is the number of occurrences of $y_i$ in the training set. To facilitate graph construction, we binarize the correlation probability $P(y_j \,|\, y_i)$. This entails converting the probability values into binary values which indicates whether a correlation exists (or not) between two labels. The operation can be written as:

$$\mathcal{E}_{ij} = \begin{cases} 0, \text{if } P(y_j \,|\, y_i) < \lambda \\ 1, \text{if } P(y_j \,|\, y_i) \geq \lambda, \end{cases} \quad (10)$$

where $\mathcal{E}$ is the binary correlation matrix that is used to form the edge set, and $\lambda$ is the hyper-parameter threshold to filter the noise edges. In our experiment, $\lambda = 1$, which means that a edge is formed when the two labels in each pair always appear together.

To encode the graph $\mathcal{G}$, we first generate the initial node features using the ICD full descriptors

for each code $y$ via Clinical-Longformer:

$$\begin{aligned} \mathrm{y} &= \{[\mathsf{CLS}], x_1, x_2, ..., x_{n-2}, [\mathsf{SEP}]\}, \\ H_v &= \mathrm{ClinicalLongformer}(\mathrm{y}), \end{aligned} \tag{11}$$

where y represents a sequence of words in the label descriptors of label $y$, $H_v \in \mathbb{R}^{n \times h_e}$, and $h_e$ is the hidden size. We use the hidden state of the first token ([CLS]) to represent the initial node feature denoted as $H^i_{\mathrm{node}}$ for the $i^{th}$ label.

With all initial node features stacked as a matrix $V = \{H^1_{\mathrm{node}}, H^2_{\mathrm{node}}, ..., H^L_{\mathrm{node}}\}$, where $V \in \mathbb{R}^{h_e \times L}$, a standard self-attention layer is then used for feature migration. To leverage the structural information, a novel spatial encoding method is used to modify the Query-Key product matrix $A^{\mathcal{G}}$ in the self-attention layer:

$$A^{\mathcal{G}}_{ij} = \frac{(H^i_{\mathrm{node}} W^{\mathcal{G}}_Q)(H^j_{\mathrm{node}} W^{\mathcal{G}}_K)^\intercal}{\sqrt{h_e}} + b_{\phi(y_i, y_j)}, \tag{12}$$

where $W^{\mathcal{G}}_Q$ and $W^{\mathcal{G}}_K$ are layer-specific weight matrices, and $\phi(y_i, y_j)$ is the spatial relation between $y_i$ and $y_j$ in graph $\mathcal{G}$, and the function $\phi(\cdot)$ is defined as the connectivity between the nodes in $\mathcal{G}$, which is the co-occurrence relation among labels. $b_{\phi(y_i,y_j)}$ is a learnable scalar indexed by $\phi(y_i, y_j)$, and shared across all layers. The attention score $A^{\mathcal{G}}_{ij}$, then, has been used to aggregate the multi-head attention for the final output:

$$h^{l+1} = \mathrm{MHA}(\mathrm{LN}(h^l)) + h^l, \tag{13}$$

where LN denotes the layer normalization, MHA denotes the multi-head self-attention, $h^l$ and $h^{l+1} \in \mathbb{R}^{L \times h_e}$ indicate the node representation of the $l^{th}$ and $(l+1)^{th}$ layers. We use the last layer to represent the label feature denoted as $H_L$. For more details on the full structure of Graphormer, please refer to the original paper (Ying et al., 2021).

**Contrastive Learning for Re-ranking** Now, we construct a code co-occurrence guided contrastive learning framework. Unlike supervised learning that aims to understand "what is what", contrastive learning adopts a different perspective by learning "what is similar or dissimilar to what". In our problem setting, we focus on the distances between a clinical document and its associated ICD codes, rather than solely between samples themselves. We consider the ground truth labels as positive samples, while the negative samples comprise all the other labels within the label space. Given $H_{\mathrm{T}}$, the representation for a clinical note $d$, and the set of representations of its corresponding ICD codes denoted as $H^+_{\mathrm{L}}$, we denote the representations of $N$ negative ICD codes randomly chosen from the ICD codes of the documents in the batch (batch size is $N$), which are not ICD codes of document $d$, as $H^-_{\mathrm{L}}$. Contrastive learning aims to learn the effective representations by pulling $d$ and $H^+_{\mathrm{L}}$ together while pushing apart $d$ and $H^-_{\mathrm{L}}$, represented as $S$ and $D$, respectively, in the equation below. The contrastive loss can be defined as:

$$\mathcal{L} = -\log \frac{S/\tau}{S/\tau + D/\tau}, \tag{14}$$

where $S = \exp(\sum_{c \in L^+_{\mathrm{L}}} \cos(H_{\mathrm{T}}, c)/|H^+_{\mathrm{L}}|)$, $D = \exp(\sum_{c' \in L^-_{\mathrm{L}}} \cos(H_{\mathrm{T}}, c')/N)$, and $\tau$ is the temperature hyper-parameter. During inference, a comparison is conducted by measuring the distance between the query clinical note and ICD codes in the embedding space, which ranks the ICD codes and then provides recommendations of the potential ICD candidates.

## 4 Experiments

### 4.1 Dataset and Pre-processing

We conduct our experiments on the publicly available benchmark MIMIC-III (Johnson et al., 2016) dataset that contains a variety of patient data types, including discharge summaries, demographic details, interventions, laboratory results, physiologic measures, and medication information. Following previous work, we are interested in the de-identified discharge summaries with annotated ICD-9 codes. There are 52,722 discharge summaries and 8,922 unique ICD-9 codes in the dataset. We mainly use three major data resources from the dataset: (1) de-identified discharge summaries (from the NOTEEVENTS table); (2) ICD-9 codes (from DIAGNOSES_ICD and PROCEDURES_ICD tables); and (3) auxiliary knowledge including DRG codes, CPT codes and drug prescriptions (from DRGCODES, CPTEVENTS, and PRESCRIPTIONS tables).

To preprocess the clinical notes, we first remove all de-identified information, then replace punctuation and atypical alphanumerical character combinations (e.g., '3a', '4kg') with white space, and lowercase every token. We truncate the discharge summaries at a maximum length of 4000 tokens. We follow Mullenbach et al. (2018) to form two settings: full codes (MIMIC-III-full) and top-50

frequent codes (MIMIC-III-top 50). In MIMIC-III-full, there are 47,719 discharge summaries for training, with 1,632 for validation, and with 3,372 for testing.

### 4.2 Implementation and Evaluation

We implement our model in PyTorch (Paszke et al., 2019) on a single NVIDIA A100 40G GPU. We use the Adam optimizer and early stopping strategies using the Micro-F1 score over the validation set as the stopping criterion to avoid overfitting. We set the initial learning rate as 5e-5 with batch size 16. We choose a learning rate scheduler which is warmed up with cosine decay, and the warm up ratio is set to 0.1. Our code is available at https://github.com/xdwang0726/ICD-contrastive-curriculum.

For evaluating the performance of our proposed model, we employ three commonly used metrics: F1-score (Micro and Macro), AUC (Micro and Macro), and precision at $K$ (P@K).

## 5 Results and Discussion

In order to asses the efficacy of our proposed framework, we compare with the existing state-of-the-art (SOTA) models, as outlined in Table 1. The top score for each metric is denoted in bold. As shown, our model outperforms in the majority of evaluation metrics, with the exception of Macro-AUC and Macro-F1 on MIMIC-III-full and MIMIC-III-top 50. Notably, our model achieves comparable performance on Micro-F1 and Micro-AUC, and improves precision at $K$ on both MIMIC-III-full and MIMIC-III-top 50. These results provide solid evidence to validate the efficacy of integrating auxiliary knowledge in the retrieval stage and leveraging code co-occurrence guided contrastive learning in the re-ranking stage.

As the occurrence frequencies of the ICD codes are imbalanced, our focus lies in assessing the efficacy of our model, specifically on the infrequently appearing ICD codes. We categorize the ICD codes into four groups based on their occurrences in the training set: [0, 10), [10, 50), [50, 500), and [500,). Figure 3 illustrates the distribution of ICD codes and their occurrence percentages across the four categorized groups in the training set, which show that the distribution of ICD frequency is highly biased, conforming to a long-tail distribution. Figures 3b and 3c present the performance of our model on MIMIC-III-full in comparison to the CAML baseline (Mullenbach et al., 2018) across the four ICD groups on Macro-AUC and Micro-F1, respectively. Our model demonstrates significant improvements for both frequent and infrequent labels on both metrics.

To confirm the specific contributions of these modules in terms of enhancing both the effectiveness and robustness of the model, we conduct ablation studies with three different settings: (a) we examine the effectiveness of using auxiliary knowledge in the retrieval stage by removing the retrieval stage and rank the ICD codes on the whole label set; (b) we examine the influence of different embedding methods by replacing the Clinical-Longformer with Clinical-BERT; and (c) we test the effectiveness of label embedding by replacing the encoding of the label with the average of word embeddings in the label descriptors. The experimental results are shown in Table 2. We also conduct case studies to qualitatively understand the effects of incorporating the label co-occurrence and the auxiliary knowledge. Two case studies have been presented in Appendix A.

**Effectiveness of Using Auxiliary Knowledge for Retrieval** We employ three distinct types of auxiliary knowledge in the retrieval stage: DRG codes, CPT codes, and drug prescriptions. As shown in Table 2, removing auxiliary knowledge leads to a decline in performance, indicating the pivotal role of the retrieval stage. This outcome further provides evidence that external knowledge effectively addresses the challenge presented by a large pool of potential ICD codes. Through integrating external knowledge, the retrieval stage attains the capability to refine the candidate list using the co-occurrence relationships between ICD codes and the auxiliary knowledge, thereby amplifying both the efficiency and accuracy of the re-ranking stage. The selection of an appropriate candidate list for a given medical record hinges upon a hyperparameter, specifically the threshold $\eta$ governing the co-occurrence between auxiliary knowledge and ICD codes. The choice of $\eta$ determines the candidate numbers that implicitly affect the overall performance of the model. Setting $\eta = 0.005$, the candidate list guarantees inclusion of 99.22% of the gold-standard ICD codes, resulting in an average of 1,460 codes in the subset. Notably, this accounts for approximately one-sixth of the complete code set. A further reduction using BM25 limits the candidate list to 1,299 on average.

| Models | MIMIC-III-full | | | | | | MIMIC-III-top 50 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | | F1 | | P@K | | AUC | | F1 | | P@5 |
| | Macro | Micro | Macro | Micro | P@8 | P@15 | Macro | Micro | Macro | Micro | |
| CAML (Mullenbach et al., 2018) | 0.895 | 0.986 | 0.088 | 0.539 | 0.709 | 0.561 | 0.875 | 0.909 | 0.532 | 0.614 | 0.609 |
| DR-CAML (Mullenbach et al., 2018) | 0.897 | 0.985 | 0.086 | 0.529 | 0.690 | 0.548 | 0.884 | 0.916 | 0.576 | 0.633 | 0.618 |
| MultiResCNN (Li and Yu, 2020) | 0.910 | 0.986 | 0.085 | 0.552 | 0.734 | 0.584 | 0.899 | 0.928 | 0.606 | 0.670 | 0.641 |
| LAAT (Vu et al., 2020) | 0.919 | 0.988 | 0.099 | 0.575 | 0.738 | 0.591 | 0.925 | 0.946 | 0.666 | 0.715 | 0.675 |
| Joint-LAAT (Vu et al., 2020) | 0.921 | 0.988 | 0.107 | 0.575 | 0.735 | 0.590 | 0.925 | 0.946 | 0.661 | 0.716 | 0.671 |
| EffectiveCAN (Liu et al., 2021) | 0.915 | 0.988 | 0.106 | 0.589 | 0.758 | 0.606 | 0.915 | 0.938 | 0.644 | 0.702 | 0.656 |
| MSMN (Yuan et al., 2022) | **0.950** | 0.992 | 0.103 | 0.584 | 0.752 | 0.599 | **0.928** | 0.947 | 0.683 | 0.725 | 0.680 |
| KEPTLongformer (Yang et al., 2022) | - | - | **0.118** | 0.599 | 0.771 | 0.615 | 0.926 | 0.947 | **0.689** | 0.728 | 0.672 |
| Ours | 0.949 | **0.995** | 0.114 | **0.603** | **0.775** | **0.623** | 0.927 | **0.947** | 0.687 | **0.732** | **0.685** |

Table 1: Comparison to previous methods across three main evaluation metrics MIMIC-III dataset. Bold: the optimal values.

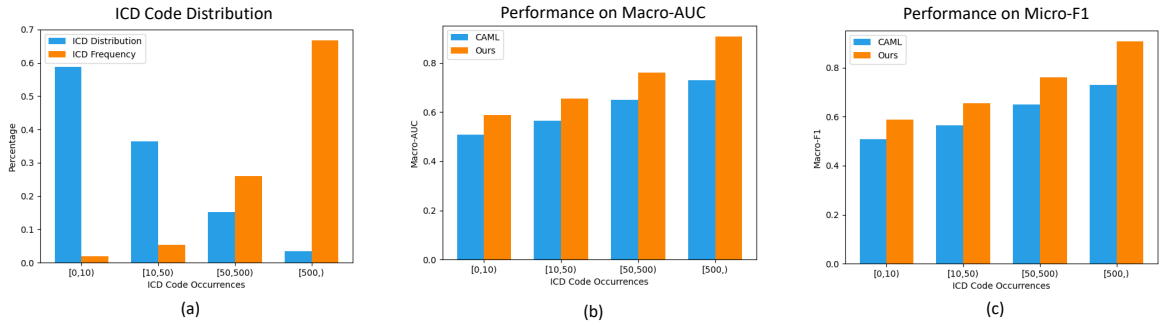


Figure 3: (a) ICD code distribution. (b) Macro-AUC performance comparison of our model and CAML on ICD codes at different frequency. (c) Micro-F1 performance comparison of our model and CAML on ICD codes at different frequency.

| Methods | F1 | | P@K | |
|---|---|---|---|---|
| | Macro | Micro | P@8 | P@15 |
| Full Model | **0.114** | **0.603** | **0.775** | **0.623** |
| w/o auxiliary knowledge | 0.097 | 0.579 | 0.748 | 0.587 |
| embedded w/ Clinical-BERT | 0.083 | 0.548 | 0.711 | 0.546 |
| w/o Graphormer | 0.102 | 0.583 | 0.753 | 0.591 |

Table 2: Ablation experiment results on the MIMIC-III-full. Bold: the optimal values.

**Comparison of Clinical-Longformer and Clinical-BERT** Increasing the maximum token limit is important in the context of clinical notes analysis as clinical texts are lengthy. Specially, in the MIMIC-III dataset, the average length of the discharge summaries is 1,596. Given this substantial token volume in the clinical notes, encoding a maximum number of tokens prior to downstream analysis becomes a pivotal requirement, which facilitates robust and meaningful subsequent analysis. To test the effectiveness of using longer sequences, we compare the model performance of Clinical-Longformer and a BERT-based pre-trained language model (i.e., Clinical-BERT) which can encode a maximum of 512 tokens. As shown in Table 2, Clinical-Longformer substantially outperforms Clinical-BERT, indicating the importance of the maximum token limit on language models in the automatic medical coding task.

**Effectiveness of Learning Label Features Using Code Co-occurrence** The graph structure has been shown to be effective in modeling code correlations and Graphormer efficiently learns code representations. The findings presented in Table 2 highlight the affirmative impact of integrating code co-occurrence into label representations. By using Graphormer, the model effectively captures and exploits the intricate connections and interdependencies among the labels, thereby improving the overall performance. This indicates that incorporating code co-occurrence information with Graphormer empowers the model to gain insights from the collaborative behaviours of the labels, consequently facilitating a more holistic comprehension of the underlying label co-relations.

## 6 Conclusion

In this paper, we regard the medical coding task as a recommendation problem and present a novel

multi-stage retrieve and re-rank framework. The primary objective of the proposed framework is twofold: to construct a curated list of ICD codes and, subsequently, to further refine the candidate list for a given medical record. Specifically, we first conduct a two-step retrieval process, incorporating auxiliary knowledge and the BM25 algorithm. This approach retrieves a concise subset of the candidate list, mitigating the challenges of a very large and imbalanced label distribution. We then use a re-ranking model to refine the previously obtained candidate list, employing code co-occurrence guided contrastive learning. Experimental results demonstrate that our proposed framework outperforms the previous SOTA, which suggests that it provides more precise and contextually grounded ICD recommendations for the given medical records. In the future, our proposed framework may be extended with more external knowledge such as the Unified Medical Language System (UMLS) and code synonymy.

## Limitations

Our usage of auxiliary knowledge is limited to external knowledge that includes DRG codes, CPT codes, and drug prescriptions, only. Other knowledge including disease-symptom, disease-lab relations, Unified Medical Language System (UMLS), and others, could also be potentially useful for the auto ICD coding task. We also acknowledge that the auxiliary knowledge we used is labeled by human annotators, which may require some extra effort. We are not quite sure about the workload for annotating different code terminologies, but we believe linking different code terminologies is important.

Our study is constrained by its evaluation limited to MIMIC-III-full and MIMIC-III-top 50 datasets, primarily concentrated on common disease. To comprehensively assess the model's performance on rare diseases, future work could benefit from a curated list of rare diseases validated by domain experts.

## Ethics Statement

We are using the publicly available clinical dataset MIMIC-III, which contains de-identified patient information. We do not see any ethics issue here in this paper.


## Acknowledgements

We would like to thank all reviewers for their comments, which helped improve this paper considerably. Computational resources used in preparing this research were provided, in part, by Compute Ontario[6], Digital Research Alliance of Canada[7], the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute[8]. This research is partially funded by The Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant to R. E. Mercer. F. Rudzicz is supported by a CIFAR Chair in AI.


## References


Tian Bai and Slobodan Vucetic. 2019. Improving medical code prediction from clinical text via incorporating online knowledge sources. *The World Wide Web Conference*, page 72–82.

Tal Baumel, Jumana Nassour-Kassis, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes a case study on ICD code assignment. In *The Workshops of the Thirty-Second AAAI Conf. on Art. Intell.*, pages 409–416.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Hypercore: Hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 3105–3114.

Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2020. Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 73–78.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th Int. Conference on Learning Representations*.


[6] https://www.computeontario.ca
[7] https://ccdb.alliancecan.ca
[8] https://www.vectorinstitute.ai/partners

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8180–8187.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proc. of the 2018 Conf. of the North American Chapter of the Ass. for Comp. Ling.: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 232–241.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards automated ICD coding using deep learning. *ArXiv*, abs/1711.04075.

Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. 2023. A review on deep neural networks for icd coding. *IEEE Trans. on Knowledge and Data Eng.*, 35(5):4357–4375.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of the 31st Int. Conference on Neural Information Processing Systems*, page 6000–6010.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. Main track.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proc. of the 60th Annual Meeting of the Assoc. for Computational Linguistics (Volume 1: Long Papers)*, pages 7109–7119.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated ICD coding. In *Proc. of the 56th Annual Meeting of the Ass. for Comp. Ling. (Vol. 1: Long Papers)*, pages 1066–1076.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proc. of the 28th ACM Int. Conf. on Information and Knowledge Management*, page 649–658.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. In *Findings of the Assoc. for Computational Linguistics: EMNLP 2022*, pages 1767–1781.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proc. of the 60th Annual Meeting of the Assoc. for Comp. Ling. (Vol. 2: Short Papers)*, pages 808–814.

## A Case Studies

We conducted case studies to qualitatively explore the impacts of integrating label co-occurrence (illustrated in Figure 4) and auxiliary knowledge (depicted in Figure 5). We compared the full model with models that did not integrate the label co-occurence and the external knowledge on the predictions of two patient records. For each patient, we present the discharge summary, ground truth ICD codes, label co-occurrence information, and auxiliary knowledge information, along with the predicted ICD codes from the full model and ablated models.

In Case 1, the ground truth ICD codes '785.51 Cardiogenic shock' and 'V49.86 Do not resuscitate status' are not explicitly mentioned in the discharge summary. The observed label co-occurrence

| Case 1: Effectiveness of Incorporating Label Co-occurrence | |
|---|---|
| Discharge Summary | Chief Complaint: heroin overdose<br>Major Surgical or Invasive Procedure: s/p intubation, s/p cvc placement SINGLE SUPINE AP PORTABLE CHEST RADIOGRAPH: An endotracheal tube is in optimal position terminating 3.5 cm above the carina. A nasogastric tube coils within the stomach, with the tip terminating in the distal stomach. No pneumothorax or large pleural effusions are seen. There is diffuse opacity overlying the entire right lung and major portion of the left upper lung, which likely represent diffuse pulmonary edema, ARDS or hemorrhage. No acute osseous abnormality seen.<br>IMPRESSION: Diffuse opacities in the right lung and left upper lung, likely represents pulmonary edema, ARDS or hemorrhage. ET tube in optimal position.<br><br>Brief Hospital Course:<br>History of Present Illness and MICU Course: Mr. [**Known lastname 12303**] is a 19 year old male with a history of polysubstance abuse most significant for intravenous heroin, who presented to the [**Hospital1 18**] ED for post-cardiac arrest care in the setting of an apparent heroin overdose.<br>Briefly, he was discharged from a rehab center in [**State 108**] one day prior to admission. Last night, at 3AM on [**2145-4-10**], his mother found him down with needles around. She immediately called 911 and initiated CPR. He was intubated in the field per the [**Location (un) 5700**] service ambulance record and dopamine and levofed were initiated; his pupils were reportedly fixed and dilated at that point. Patient cooling was also performed via ice packs. .....<br>In the [**Hospital1 18**] ED he was on three pressors (epinephrine, levophed, and vasopressin). His blood cases were checked twice and showed 6.79/86/61 -->6.87/67/82. He was transferred to the MICU. In the MICU, he did not have a femoral pulse. A cardiac monitor was placed and he was noted to have pulseless electrical activity. ACLS was initiated. He received sodium bicarbonate, calcium chloride, d50, NS, and boluses of epinephrine. His rhythm converted to ventricular fibrillation and he was shocked.<br>He then converted to PEA and regained a pulse after another bolus of epinephrine. The family was present. The code lasted just under ten minutes. After discussion with the family, the decision was made not to escalate care (see Dr. [**Last Name (STitle) **]??????s note).<br>He remained on three pressors with ventilatory support. Within one hour he became bradycardic and expired. See written death note in the chart. The organ bank declined the case for donation. The Medical examiner accepted the case. The family declined discretionary autopsy. Death report and other necessary documentation was filed. |
| Ground Truth ICD Codes | 427.5 Cardiac arrest; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; V49.86 Do not resuscitate status; 518.81 Acute respiratory failure; 785.51 Cardiogenic shock;<br>99.60 Cardiopulmonary resuscitation, not otherwise specified;<br>304.71 Combinations of opioid type drug with any other drug dependence, continuous |
| Examples of Label Co-occurrence Information | 1. 427.5 Cardiac arrest relates to 785.51 Cardiogenic shock<br>2. 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours relates to V49.86 Do not resuscitate status |
| Predictions of Full Model | 427.5 Cardiac arrest; 96.04 Insertion of endotracheal tube; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours;<br>965.01 Poisoning by heroin; 99.6 Cardiopulmonary resuscitation, not otherwise specified; 785.51 Cardiogenic shock; V49.86 Do not resuscitate status |
| Predictions of No Label Co-occurrence | 427.5 Cardiac arrest; 96.04 Insertion of endotracheal tube; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours;<br>965.01 Poisoning by heroin; 99.6 Cardiopulmonary resuscitation, not otherwise specified;<br>969.6 Poisoning by psychodysleptics (hallucinogens) |

Figure 4: Case study on the effectiveness of incorporating label co-occurrence. Correctly predicted labels are marked in green and the incorrect ones are marked in red.

between '427.5 Cardiac arrest' and '785.51 Cardiogenic shock', as well as co-relation between '96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours' and 'V49.86 Do not resuscitate status' provide strong indicators suggesting the presence of the codes '785.51' and 'V49.86'. Without the label co-occurrence signals, the ablated model missed the predictions of codes '785.51' and 'V49.86', indicating a failure to leverage latent label information.

In Case 2, the patient has been diagnosed with '244.9 Unspecified acquired hypothyroidism' with less explicit information in the discharge summary. Notably, the presence of the medication 'Levothyroxine' in the drug prescription, an element of auxiliary knowledge, suggests that the patient is likely to have acquired hypothyroidism. The ablated model, lacking the auxiliary knowledge, misses the prediction of code '244.9'. The aforementioned Cases 1 and 2 highlight the benefits of incorporating label co-occurrence and auxiliary knowledge, respectively.

| Case 2: Effectiveness of Incorporating Auxiliary Knowledge | |
|---|---|
| Discharge Summary | Chief Complaint: Subdural hematoma<br>History of Present Illness: 78 year-old male with hypertension, ITP on Rituximab transferred from OSH for further management of SDH. Felt poorly yesterday. Woke up this morning with severe HA. Unresponsive in EMS. Went to [**Hospital1 **], found to have decerebrate posturing, fixed and dilated pupils. CT head with large left-sided SDH with 2mm shift, and transtorial herniation. Intubated (succ/etomidate), mannitol. Also received atropine for unknown reason. ... Discussed with neurosurgery, radiology; determined to benefit in intervention at this point. Per report from ED resident, patient converted to CMO, and awaiting arrival of family prior to extubation. Propofol restarted for comfort. On transfer to ICU, 67, 151/65, 10, 100% AC 10/500 PEEP 5, FiO2 100%. On the floor, patient is intubated and not responsive.<br>Brief Hospital Course: 78M with hypertension, ITP with subdural hematoma complicated by mass effect. Expired shortly after admission.<br>#. Subdural hematoma: In context of thrombocytopenia and known hypertension. Complicated by mass effect. Patient noted intially to be decorticate. Unresponsive with fixed/dilated pupils off of sedation. With down titrating ventilatory support, patient with rare breaths and with low tidal volumes. Discussed with family; plan for comfort.<br>#. ITP: Thrombocytopenic. Held off on platelet transfusion as would not change outcome.<br>#. Hypertension: Held anti-hypertensives. |
| Ground Truth ICD Codes | V58.65 Long-term (current) use of steroids; 401.9 Unspecified essential hypertension; 432.1 Subdural hemorrhage; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; 244.9 Unspecified acquired hypothyroidism; 348.4 Compression of brain; 287.31 Immune thrombocytopenic purpura |
| Examples of Using Auxiliary Knowledge | 1. Levothyroxine relates to 244.9 Unspecified acquired hypothyroidism |
| Predictions of Full Model | 287.31 Immune thrombocytopenic purpura; 348.4 Compression of brain; 348.5 Cerebral edema; 401.9 Unspecified essential hypertension; 432.1 Subdural hemorrhage; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; 244.9 Unspecified acquired hypothyroidism; E888.9 Unspecified fall; V66.7 Encounter for palliative care |
| Predictions of No Auxiliary Knowledge | 287.31 Immune thrombocytopenic purpura; 348.4 Compression of brain; 348.5 Cerebral edema; 401.9 Unspecified essential hypertension; 432.1 Subdural hemorrhage; 96.71 Continuous invasive mechanical ventilation for less than 96 consecutive hours; E888.9 Unspecified fall; V66.7 Encounter for palliative care; 852.20 Subdural hemorrhage following injury without mention of open intracranial wound, unspecified state of consciousness |

Figure 5: Case study on the effectiveness of incorporating auxiliary knowledge. Correctly predicted labels are marked in green and the incorrect ones are marked in red.