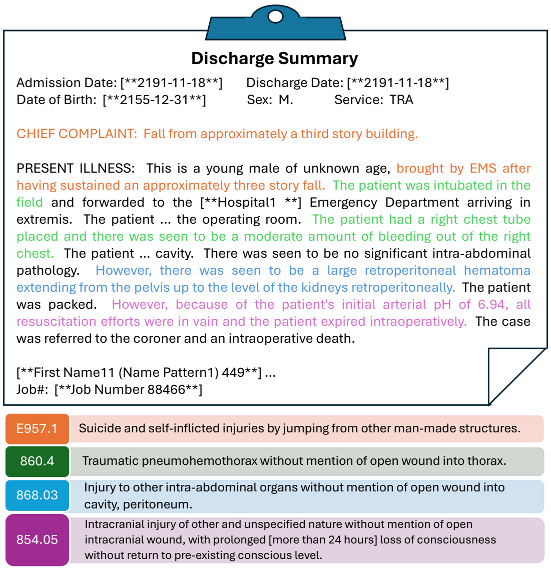
Structured Information Matters:
Explainable ICD Coding with Patient-Level Knowledge Graphs

Anonymous ACL submission

Abstract

Mapping clinical documents to standardised clinical vocabularies is an important task, as it provides structured data for information retrieval and analysis, which is essential to clinical research, hospital administration and improving patient care. However, manual coding is both difficult and time-consuming, making it impractical at scale. Automated coding can potentially alleviate this burden, improving the availability and accuracy of structured clinical data. The task is difficult to automate, as it requires mapping to high-dimensional and long-tailed target spaces, such as the International Classification of Diseases (ICD). While external knowledge sources have been readily utilised to enhance output code representation, the use of external resources for representing the input documents has been underexplored. In this work, we compute a structured representation of the input documents, making use of document-level knowledge graphs (KGs) that provide a comprehensive structured view of a patient's condition. The resulting knowledge graph efficiently represents the patient-centred input documents with 23% of the original text while retaining 90% of the information. We assess the effectiveness of this graph for automated ICD-9 coding by integrating it into the state-of-the-art ICD coding architecture PLM-ICD. Our experiments yield improved Macro-F1 scores by up to 3.20% on popular benchmarks, while improving training efficiency. We attribute this improvement to different types of entities and relationships in the KG, and demonstrate the improved explainability potential of the approach over the text-only baseline.



Discharge Summary

Admission Date: [**2191-11-18**] Discharge Date: [**2191-11-18**]
Date of Birth: [**2155-12-31**] Sex: M. Service: TRA

CHIEF COMPLAINT: Fall from approximately a third story building.

PRESENT ILLNESS: This is a young male of unknown age, brought by EMS after having sustained an approximately three story fall. The patient was intubated in the field and forwarded to the [**Hospital1 **] Emergency Department arriving in extremis. The patient ... the operating room. The patient had a right chest tube placed and there was seen to be a moderate amount of bleeding out of the right chest. The patient ... cavity. There was seen to be no significant intra-abdominal pathology. However, there was seen to be a large retroperitoneal hematoma extending from the pelvis up to the level of the kidneys retroperitoneally. The patient was packed. However, because of the patient's initial arterial pH of 6.94, all resuscitation efforts were in vain and the patient expired intraoperatively. The case was referred to the coroner and an intraoperative death.

[**First Name11 (Name Pattern1) 449**] ...
Job#: [**Job Number 88466**]

E957.1	Suicide and self-inflicted injuries by jumping from other man-made structures.
860.4	Traumatic pneumohemothorax without mention of open wound into thorax.
868.03	Injury to other intra-abdominal organs without mention of open wound into cavity, peritoneum.
854.05	Intracranial injury of other and unspecified nature without mention of open intracranial wound, with prolonged [more than 24 hours] loss of consciousness without return to pre-existing conscious level.

Figure 1: Example of ICD Coding over MIMIC-III. The discharge summary (HADM ID: 104128) is annotated with four ICD codes.

1 Introduction

Clinical coding is the process of allocating standardized codes to diagnoses, treatments, procedures, and medical services detailed in patient electronic records or paper notes. This multi-label classification task offers advantages across various domains, including audit procedures, decision support systems and medical billing (Blundell, 2023). Various coding systems are designed to encode specific information within patient records. Our work focuses on the International Classification of Diseases (ICD-9) (Organization et al., 1978), a widely recognized coding system that holds a pivotal role in encoding diagnostic and procedural information. This process is commonly known as ICD coding. An example is shown in Figure 1.

Manual code assignment is typically costly, labor-intensive, and error-prone (Nguyen et al., 2018). In recent years, automated clinical coding, powered by cutting-edge deep learning techniques, has significantly advanced the field, improving accuracy, increasing efficiency, and reducing overall costs (Ji et al., 2022; Teng et al., 2022).

The main challenge in clinical coding arises from the extremely imbalanced distribution of the label space. For instance, in the case of MIMIC-III

(Johnson et al., 2016), there are 8,692 unique ICD-9 codes, of which 4,115 codes (47.3%) occur fewer than 6 times (Yang et al., 2022). Considering this long-tailed distribution of codes, previous work has explored integrating diverse external knowledge to enhance the representation of codes and patients. Among these external knowledge sources, knowledge graphs play an important role in improving the performance of ICD coding by providing not only semantic information but also structured information. However, most research focuses on representing ICD codes through various graphs that are built based on these codes themselves (Rios and Kavuluru, 2018; Xie et al., 2019; Cao et al., 2020; Lu et al., 2020; Song et al., 2021; Michalopoulos et al., 2022). Efforts to construct patient-level knowledge graphs remain largely underexplored in both ICD coding and the broader clinical domain.

The patient-level knowledge graph offers an intuitive representation and visualization of a patient's clinical condition, providing healthcare professionals with valuable insights. Meaningful causal relationships between entities, such as symptoms that support a diagnosis, tests performed, and treatments derived from these findings, enable patient-level knowledge graphs to facilitate more efficient decision-making for physicians and medical staff. However, critical questions remain unanswered: what elements should constitute a patient's knowledge graph, including problems, symptoms, tests, treatments, drugs, dosages, and frequencies? And how to evaluate the quality of such graphs and assess their utility and impact on tasks such as patient-level classification and explainability?

To the best of our knowledge, Yuan et al. (2021) is the only work which proposes a medical graph specifically designed for individual patients in ICD coding task. The graph integrates a disease hierarchy based on ICD-10 and a causal graph of diseases. Entities in the causal graph, including symptoms, signs, and diseases are identified from documents using the named-entity recognition (NER) technique. The model also leverages GCN to represent the nodes in the graph. It enhances the patient representation by integrating it with the raw clinical text and patient information. However, it does not cover a wide range of entity categories and capture the diverse relationships among them, which can provide a more comprehensive understanding about a patient's medical history. Additionally, this work lacks a systematic evaluation of graph quality and an analysis of the determination of its constituent components.

To close these gaps, we construct patient-level knowledge graphs that provide a wide range of entity types and relationships. This comprehensive graph offers explicit context to a patient's situation, by providing diagnostic, posology, anatomical and the temporal information of clinical events identified in the patient records. We integrate this patient-level knowledge graph into the state-of-the-art ICD coding architecture, PLM-ICD (Huang et al., 2022), demonstrating improved coding performance.

The contributions of this work are:

(i) We develop a comprehensive patient-level knowledge graph encompassing a wide coverage of 14 distinct entity types connected by five types of relationships. We evaluate the informativeness of the graph by measuring the information loss relative to the patient notes from which the graph is retrieved. Our results demonstrate that the knowledge graph effectively distills essential information from patient notes into a more concise and structured format, achieving a significant reduction in size—extracting only 23% of the original content—while retaining 90% of the information. This represents a ***Statistical Perspective*** for evaluating the quality of the graph.

(ii) We conducted experiments to assess the effectiveness of integrating graph representations into ICD coding. The results demonstrate that the additional structured information provided by the graph enhances coding performance, improving the F1-score by 1.36% compared to the base model. This improvement is significant for this developed task. This also serves as an evaluation of the patient-level knowledge graph from a ***Representational Perspective***, capturing both semantic and structural information.

(iii) We address the question of '***What elements should constitute a patient's knowledge graph?***' through an ablation study from two evaluation perspectives. We analyse the impact of various types of entities and relationships on the information retaining and coding performance.

(iv) We perform a case study and showcase the model's ability to offer high-quality explanations by providing accurate and concise evidence which supports the model's prediction.

2 Related Work

Architecture Over the past decade, the field of clinical coding has witnessed significant advancements, evolving from traditional rule-based methods (Pereira et al., 2006; Crammer et al., 2007) to advanced machine learning and deep learning approaches. Researchers have recently explored the application of cutting-edge NLP techniques, including attention mechanisms and transformer models. The architecture of these models has become increasingly sophisticated, with common architecture incorporating CNN-based (Mullenbach et al., 2018), LSTM-based (Catling et al., 2018), and transformer-based encoders (Zhang et al., 2020; Chalkidis et al., 2020; Ji et al., 2021), often paired with label-wise attention layers (Vu et al., 2020; Sun et al., 2021; Dong et al., 2021; Liu et al., 2021; Van Aken et al., 2022). Recent studies also highlight the challenge of efficiently applying transformer models to represent the inherently lengthy clinical documents. These approaches leverage transformers handling long sequences, notably Longformer (Yang et al., 2022) and BigBird (Michalopoulos et al., 2022).

External Knowledge Representations A major challenge in this field is classifying within a large target space, where the distribution of codes is highly uneven, commonly described as a 'big-head long-tail' distribution. This imbalance hinders the model's effectiveness in recognising patterns associated with categories with few samples. To address this issue, researchers have turned to external knowledge to enhance the representations of both patients and codes. For patient representation, this includes data augmentation (Falis et al., 2022; Song et al., 2021) and knowledge graphs (Yuan et al., 2021). In terms of code representation, external knowledge is drawn from code descriptions (Feucht et al., 2021), synonyms (Yuan et al., 2022), relevant documents (Wang et al., 2022), code hierarchy (Falis et al., 2019; Yang et al., 2022), synthetic data (Falis et al., 2022), and knowledge graphs.

Knowledge Graph in ICD Coding Rios and Kavuluru (2018) represents ICD codes using their hierarchical structure, applying two layers of graph convolutional networks (GCN) to leverage this structured knowledge. Song et al. (2021) improves this model by replacing the GCN with graph gated recurrent neural networks (GRNN) (Li et al., 2015). Cao et al. (2020) introduces Co-Graph, which models co-occurrence correlations between codes. This graph is represented by its adjacency matrix and GCN. Lu et al. (2020) constructs three types of graphs: a label hierarchy graph of class taxonomy, a semantic similarity graph derived from code descriptions, and a code co-occurrence graph similar to the approach in Cao et al. (2020). Michalopoulos et al. (2022) establishes connections between codes using normalized point-wise mutual information and also employs GCN to capture the representations of codes from this graph.

3 Methodology

In this section, we detail the construction of patient-level knowledge graphs and their integration into the PLM-ICD coding architecture.

Patient-Level Knowledge Graph Construction We aim to construct patient-level knowledge graphs that comprehensively represent a patient's medical history, encompassing diseases, treatments, tests, drugs, dosages, frequencies, strengths, and so on, as well as the relationships between these entities. We employ named-entity recognition (NER) and relation extraction (RE) models provided by Healthcare NLP library (John Snow Labs, 2024) to extract these concepts.

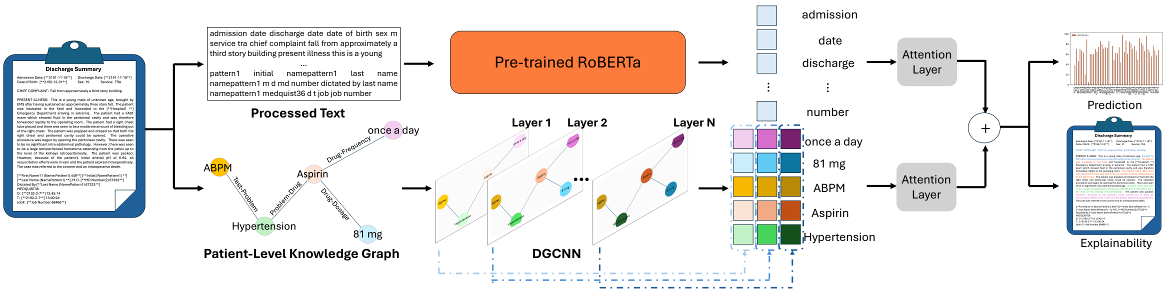
Out of the available RE models in Healthcare NLP, we select five models based on the quantity of triples extracted and their uniformity across all documents. The selected RE models are (ordered by frequency) *'Clinical Relationship'* (CR), *'Temporal Events'* (TE), *'Posology Relationship'* (PR), *'Bodypart-Directions'* (BD) and *'Bodypart-Problem'* (BP). These models collectively identify 14 different types of entities. Detailed information about model selection, selected RE models and statistics of the extracted entities and relationships can be found in in Appendix A.1.

The output of these relationship extraction (RE) models includes two identified entities, their respective types, and the relationship between them. When constructing a patient's knowledge graph, we represent this information as triples in the format $< \ entity1, relationship, entity2 \ >$ (e.g., $< lisinopril, drug\text{-}strength, 40mg >$). The resulting patient-level knowledge graphs contain four types of information (For a visualisation, consult Appendix A.2):

Diagnostic Information (CR): Revealing the interrelationships among problems, treatments, and tests;

Temporal Information (TE): Capturing the sequence of clinical events;

Posology Information (PR): Providing details on drug regimens, including dosage, duration,



Discharge Summary

Admission Date: [**2191-11-18**] Discharge Date: [**2191-11-18**]
Date of Birth: [**2155-12-31**] Sex: M. Service: TRA

CHIEF COMPLAINT: Fall from approximately a third story building.

PRESENT ILLNESS: This is a young male of unknown age, brought by EMS after having sustained an approximately three story fall. The patient was intubated in the field and forwarded to the [**Hospital1 **] Emergency Department arriving in extremis. The patient had a FAST exam which showed fluid in the peritoneal cavity and was therefore forwarded rapidly to the operating room. The patient had a right chest tube placed and there was seen to be a moderate amount of bleeding out of the right chest. The patient was prepped and draped so that both the right chest and peritoneal cavity could be opened. The operative procedure was begun by opening the peritoneal cavity. There was seen to be no significant intra-abdominal pathology. However, there was seen to be a large retroperitoneal hematoma extending from the pelvis up to the level of the kidneys retroperitoneally. The patient was packed. However, because of the patient's initial arterial pH of 6.94, all resuscitation efforts were in vain and the patient expired intraoperatively. The case was referred to the coroner and an intraoperative death.

[**First Name11 (Name Pattern1) 449**] [**Initial (NamePattern1) **]
[**Last Name (NamePattern1) **], M.D. [**MD Number(2) 67332**]
Dictated By:[**Last Name (NamePattern1) 67333**]
MEDQUIST36
D: [**2192-2-7**] 12:45:14
T: [**2192-2-7**] 13:00:24
Job#: [**Job Number 88466**]

admission date discharge date date of birth sex m service tra chief complaint fall from approximately a third story building present illness this is a young
...
pattern1 initial namepattern1 last name namepattern1 m d md number dictated by last name namepattern1 medquist36 d t job job number

Figure 2: Architecture of the proposed model. The processed discharge summary as input is encoded using a pre-trained RoBERTa, while its corresponding patient-level knowledge graph inputs a DGCNN module, with final representations obtained by concatenating node features from all layers. Both representations are fed into separate label-wise attention layers, after which the weighted outputs are concatenated, using for ICD code prediction.

Split	Avg $\lvert T\rvert$	Avg $\lvert N\rvert$	Avg T in N	Min/Max T	Min/Max N
Full	1513.5	183.0	342.3	0/1954	0/903
Top-50	1612.0	196.8	366.8	6/1689	3/774

Table 1: Statistics of nodes and tokens per processed document in MIMIC-III datasets. T stands for tokens, N stands for graph nodes. *'Avg'* represents averages over all documents.

Dataset	Text Entropy	Graph Entropy	Ratio (%)
Full	8.33	7.48	89.95
Top-50	8.41	7.61	90.52

Table 2: The Information entropy of processed text and serialised graph. The *'Ratio'* measures how much information is retained.

strength, and frequency, as well as their interrelationships;

Anatomical Information (BD and BP): Illustrating the connections between problem or directions and specific body parts.

The statistics of the graphs extracted from the two MIMIC-III datasets, Full and Top-50, are summarized in Table 1. On average, the graphs contain approximately 190 nodes, with each node typically comprising around two tokens. The largest graph in the dataset includes 903 nodes, while some documents don't have any extracted graphs.

Furthermore, we evaluate the quality of the constructed graphs from a ***Statistical Perspective*** by measuring information loss. Specifically, we calculate the average information entropy of the original text and the serialized graph. As shown in Table 2, our analysis indicates that the extracted content accounts for less than 23% of the original size, yet retains approximately 90% of the information. This highlights the efficiency of our patient-level knowledge graph in significantly compressing the text while preserving the majority of its informational content. (For details of the information entropy methodology and further results of the ablation study, conducted by removing each type of entity and relationship, please refer to Appendix A.3.)

Task Definition ICD coding is formulated as a multi-label classification task. Given a clinical document (discharge summary in MIMIC-III) of a patient, automated coding module aims to assign the correct ICD codes which represent the diseases or procedures. Specifically, we define a clinical document with N_t tokens as $\mathbf{d} = \{t_1, t_2, ..., t_{N_t}\}$. The goal is to predict a distribution of labels $\mathbf{p} = \{p_1, p_2, ..., p_{N_c}\}$, where N_c denotes the total number of codes in the label space. The final set of assigned codes is the ones that exceed a pre-defined probability threshold.

The proposed framework is shown in Figure 2. The subsequent sections will provide a detailed description of each component of the framework.

Text Embedding - Pre-trained Language Model

To embed the textual data, we utilize RoBERTa-PM (Lewis et al., 2020), a transformer model pretrained on biomedical abstract and clinical documents.

The pre-processing of the raw text in MIMIC-III datasets follows Mullenbach et al. (2018). Following PLM-ICD, we divide each document into segments of equal length of l tokens. The number of segments per document is represented as N_s and varies across different samples. Thus, each segment comprises a sequence of tokens that represent a portion of the document:

$$s_i = \{t_j | l \cdot i \leq j < l \cdot (i + 1)\}. \qquad (1)$$

The document representation $\mathbf{H_t}$ is formed by concatenating the hidden representations of each segment:

$$\mathbf{H}_{\mathrm{t}} = \text{concat}(PLM(s_1), ..., PLM(s_{N_s})), \quad (2)$$

where $PLM(s_i)$ denotes the representation for segment s_i embedded by RoBERTa-PM.

Graph Embedding - Deep Graph Convolutional Neural Network The Deep Graph Convolutional Neural Network (DGCNN) (Zhang et al., 2018) we refer to in this work is an end-to-end architecture designed for graph classification tasks. But we represent the graph using the hidden state from the final layer of DGCNN, just before the SortPooling layer in the original framework, as this configuration is found to yield the best performance based on initial experimental results.

Given a patient's knowledge graph G, we can obtain its adjacency matrix $\mathbf{A}$ and diagonal degree matrix $\mathbf{D}$. The hidden state of the first graph convolution layer is as follows:

$$\mathbf{H}_{\mathrm{g}}^{1} = f(\mathbf{D}^{-1}\mathbf{A}\mathbf{X}\mathbf{W}), \quad (3)$$

where $\mathbf{X} \in R^{N_n \times d_n}$ denotes the node representation matrix with dimension d_n; N_n represents the number of nodes in the graph; $\mathbf{W} \in R^{d_n \times d_n'}$ is a trainable parameter matrix, in which d_n' defines the dimension of code representation for the next convolution layer; f is a nonlinear activation function.

DGCNN adopts multiple convolution layers, as it allows for the extraction of multi-scale local substructure features. Therefore, the output of the m^{th} graph convolution layer is represented as follows:

$$\mathbf{H}_{\mathrm{g}}^{m+1} = f(\mathbf{D}^{-1}\mathbf{A}\mathbf{H}_{\mathrm{g}}^{m}\mathbf{W}^{m}), \quad (4)$$

where $\mathbf{H}_{\mathrm{g}}^{0} = \mathbf{X}$. The final representation of patient's knowledge graph $\mathbf{H}_{\mathrm{g}}$ is the concatenation of the features from all $[\mathbf{H}_{\mathrm{g}}^{1}, ..., \mathbf{H}_{\mathrm{g}}^{N_y}]$, where N_y is the number of graph convolution layers.

Multi-Head Label-Wise Attention To capture label-specific information and assign varying attention weights to fragments (tokens or nodes) for each label, we incorporate a label-wise attention layer following the patient representation. Instead of just feeding the concatenated representation of text $\mathbf{H}_{\mathrm{t}}$ and graph $\mathbf{H}_{\mathrm{g}}$ to a single attention layer, we utilize a multi-head attention mechanism. This approach enables the model to focus on information from different representation sub-spaces. Consequently, $\mathbf{H}_{\mathrm{t}}$ and $\mathbf{H}_{\mathrm{g}}$ are processed through separate label-wise attention layers. The attention score matrices are defined as follows:

$$\alpha_{\mathrm{t}} = \text{softmax}(\mathbf{V}_1 \tanh(\mathbf{V}_2 \mathbf{H}_{\mathrm{t}}), \quad (5)$$

$$\alpha_{\mathrm{g}} = \text{softmax}(\mathbf{V}_3 \tanh(\mathbf{V}_4 \mathbf{H}_{\mathrm{g}}), \quad (6)$$

where $\mathbf{V}_{1-4}$ are trainable linear transformation matrices. The weighted label-specific representations are calculated as follows:

$$\mathbf{Z}_{\mathrm{t}} = \mathbf{H}_{\mathrm{t}}\alpha_{\mathrm{t}}^{T}, \mathbf{Z}_{\mathrm{g}} = \mathbf{H}_{\mathrm{g}}\alpha_{\mathrm{g}}^{T}. \quad (7)$$

Finally we concatenate them to form a representation for the individual patient $\mathbf{Z} = [\mathbf{Z}_{\mathrm{t}}, \mathbf{Z}_{\mathrm{g}}]$. The probability of predicting label i is calculated by:

$$\mathbf{p}_i = \sigma(\mathbf{L}_i \cdot \mathbf{Z}_i), \quad (8)$$

where $\mathbf{L}_i$ is the representation of the i^{th} label and $\mathbf{Z}_i$ is the label-specific patient representation. The final predicted soft-maxed probability vector $\hat{\mathbf{y}}$ and true labels $\mathbf{y}$ are used to compute the binary cross-entropy loss:

$$\mathcal{L}(\mathbf{y}, \mathbf{p}) = -\frac{1}{|\mathbf{y}|}\sum_{i=1}^{|\mathbf{y}|}(\mathbf{y}_i \log \hat{\mathbf{y}}_i + (1-\mathbf{y}_i)\log(1-\hat{\mathbf{y}}_i)) . \quad (9)$$

4 Empirical Evaluation

4.1 Experiment Setup

Datasets and Metrics Like most evaluation methods for multi-label classification tasks, clinical coding is typically assessed using three standard metrics: F1, AUC and Precision@N. In this work, we utilize these metrics to evaluate the models on two commonly used datasets: MIMIC-III Full and MIMIC-III Top-50.

MIMIC-III is a publicly accessible database comprising de-identified health data from patients admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. The standard clinical coding task involves using discharge summaries from the MIMIC-III dataset to assign ICD-9 codes, which include discharge diagnoses and procedures.

The MIMIC-III Full dataset includes 52,723 documents from 41,126 patients, with each document containing a median of 1,375 words and 14 codes. The MIMIC-III Top-50 dataset focuses on the top 50 most frequent diagnosis and procedure codes from the Full dataset. It consists of 11,368 documents from 10,356 patients, with a median of 1,478 words and 5 codes per document.

	MIMIC-III Full					**MIMIC-III Top-50**				
	F1		AUC		Precision	F1		AUC		Precision
Model	Macro	Micro	Macro	Micro	P@8	Macro	Micro	Macro	Micro	P@5
MultiResCNN	9.0	55.2	91.0	98.6	73.4	59.29	66.24	89.30	92.04	61.56
2Stage	10.5	58.4	94.6	99.0	74.4	**68.93**	**71.83**	**92.58**	**94.52**	66.72
JointLAAT	10.2	57.5	92.1	98.8	73.5	66.95	70.84	92.36	94.24	66.36
MSMN	10.24	58.70	**94.78**	**99.15**	75.45	66.68	71.19	92.12	94.21	66.86
PLM-ICD	9.69	59.06	92.12	98.83	**76.72**	64.61	70.33	91.16	93.63	66.11
Our Model	**11.05**	**59.72**	92.37	98.75	76.59	67.81	71.63	92.04	94.22	**67.08**

Table 3: Results on the MIMIC-III Full and Top-50 test sets. The best results are highlighted in bold.

Implementation Details We train our model using four 80GB NVIDIA A100 GPUs within an environment configured with CUDA 11.1 and PyTorch 1.12.0. Detailed implementation hyperparameters for both our model and PLM-ICD are provided in Appendix A.4.

Baselines To demonstrate the effectiveness of our model, we compare it with five current SOTA approaches.

PLM-ICD (Huang et al., 2022), leverages transformer-based models specifically pre-trained on biomedical and clinical texts. It achieves SOTA performance on both MIMIC-III and MIMIC-IV datasets (Edin et al., 2023). We select it as our base model due to its strong performance as a widely used baseline and its simple structure, which facilitates the integration with the graph representation module.

MultiResCNN (Li and Yu, 2020) employs a multi-filter convolutional layer to capture text patterns of varying lengths and a residual convolutional layer to expand the receptive field.

2Stage (Nguyen et al., 2023) leverages the hierarchical properties of codes to perform predictions in two sequential steps.

JointLAAT (Vu et al., 2020) introduces a hierarchical joint learning mechanism to address label imbalance.

MSMN (Yuan et al., 2022) utilizes synonyms with multi-head attention mechanism, achieving another state-of-the-art performance on MIMIC-III Full.

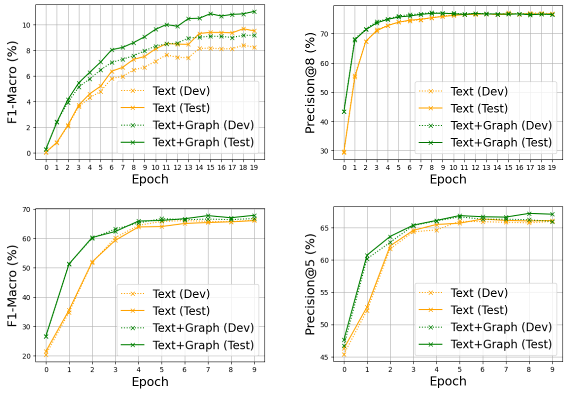


Figure 3: By-epoch performance comparison of our model and PLM-ICD by means of Macro-F1 / P@8 on MIMIC-III Full (top row) and Macro-F1 / P@5 on MIMIC-III Top-50 (bottom row).

Remove	μ**F1**	m**F1**	μ**AUC**	m**AUC**	**P@8**
Full	11.05	59.72	92.37	98.75	76.59
−BP	10.38	59.60	92.39	98.86	76.72
−PR	10.33	59.65	92.39	98.84	76.95
−TE	10.34	59.45	92.53	98.86	76.62
−CR	10.07	59.35	92.63	98.89	76.74
−BD	10.61	59.51	92.24	98.79	76.51
−drug	10.52	59.44	92.23	98.77	76.61
−problem	9.77	59.26	92.35	98.86	76.95
−treatment	10.76	59.66	92.33	98.81	76.76
−test	10.72	59.59	92.32	98.81	76.54

Table 4: Results of ablation study on the MIMIC-III Full dataset. Removing all relationships and entities of a specified type. μ and m denote Macro and Micro averages, respectively.

4.2 Quantitative Results

A. Does integrating graph-based representation enhance the ICD coding performance? This experiment aims to verify if integrating the patient-level knowledge graph benefits the representation of the patient, consequently enhances the performance of ICD coding. The results shown in Table 3 indicate that our model outperforms its base model PLM-ICD significantly on the F1-Macro score by 1.36% and 3.20% on the Full and Top-50 datasets, respectively. F1-Macro score is the primary metric for this task due to its effectiveness in balancing precision and recall across classes and its robustness in classification problems. Our model exhibits more noticeable performance improvements on frequent labels and demonstrates overall advancements across all metrics. Moreover, our model remains highly competitive compared to other state-of-the-art methods, achieving the highest F1 scores on full label set.

Additionally, our model achieves higher scores in the early epochs (see Figure 3), highlighting its efficiency when computational resources are constrained. The most significant improvements occur within the first three epochs, indicating that the structured information is efficiently captured early. These findings further validate the quality of the constructed graphs, demonstrating their effectiveness in patient representation (***Statistical Perspective***) by providing not only semantic information but also additional structured information.

B. What elements should constitute a patient's knowledge graph?

Relationship We conduct an ablation study to assess the impact of different types of relationships in the graph on patient representation. By removing a single type of relationship from the complete graph, we observe that the removal of any relationship leads to a noticeable decrease in performance. Despite this, the performance still remains superior to the base model PLM-ICD by at least 0.4% on F1-Macro score. Excluding the *'Clinical Relationship'* (CR) results in the most substantial drop in performance, indicating its critical importance in patient representation. From Table 5 in Appendix A we can see that the number of *'Clinical Relationships'* (CR) is similar to *'Temporal Events'* (TE) in MIMIC-III Full dataset. But its exclusion causes a more pronounced decline, suggesting that its significance lies not only in its quantity but also in the quality of information it provides about the patient. This is intuitive, as *'Clinical Relationships'* (CR) inherently capture the essential aspects of a patient's profile—such as medical problems, treatments, and diagnostic tests—that are directly relevant to predicting diseases and procedures codes. Conversely, *'Bodypart-Directions'* (BD) has the least impact on ICD coding.

Entity We conduct another ablation study by removing entities of the four most occurring types: *'Problem'*, *'Test'*, *'Treatment'*, and *'Drug'* (ordered by frequency). The removal of *'Problem'* has the most significant impact on the F1-Macro score, indicating that *'Problem'* plays a crucial role in the graph representation. This finding also make sense intuitively, as *'Problem'* constitutes the largest portion of the graph and is most closely related to the objective of diagnosing the patient.

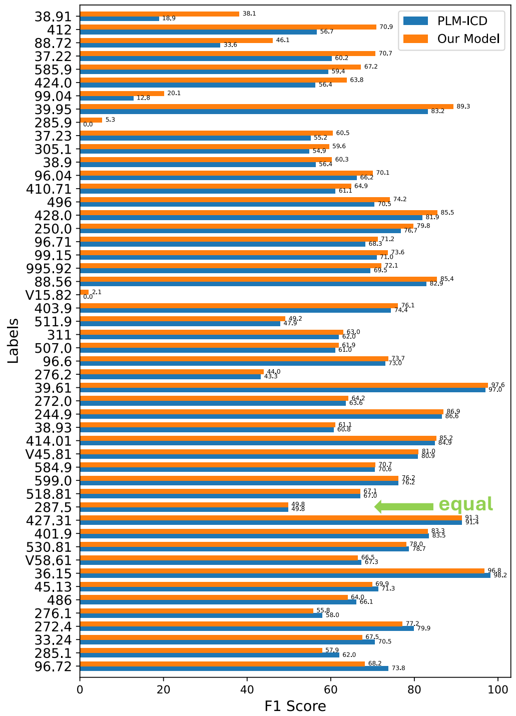


Figure 4: F1 performance comparison on each of the top-50 codes between our model and PLM-ICD, ranked by the performance difference between the two models.

4.3 Qualitative Results

C. How does the patient-level knowledge graph help the classification for specific codes? To further analyse performance at the label level, we compute the F1 scores for our model and PLM-ICD on the MIMIC-III Top-50 dataset for each code (see Figure 4). The results reveal that our model outperforms PLM-ICD on 37 codes out of 50. Notably, our model achieves scores for codes `285.9` (*Anemia, unspecified'*) and `V15.82` (*Personal history of tobacco use'*), which PLM-ICD totally fails.
To better understand how graphs enhance patient representations, we visualize the label-specific representations of all samples in the test set (see Figure 5). We focus on the codes `412` (*Old myocardial infarction'*) and `39.95` (*Hemodialysis'*) (see Appendix A.5), where both our model and PLM-ICD demonstrate good performance. This choice avoids complications from low scores, which may result in erratic embeddings that are challenging to visualize, such as the case of `38.91` *'Arterial Catheterization'*. Samples with the corresponding labels are highlighted in red. Specifically, we reduce the dimensionality of the original representations $\mathbf{Z_i}$ using t-SNE. For code `412`, our model exhibits a noticeably higher density of instances

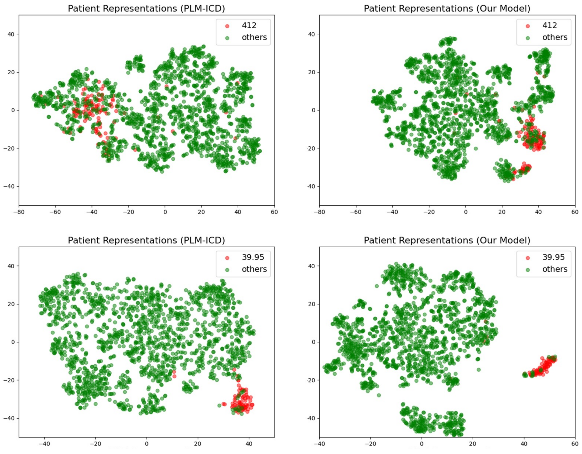
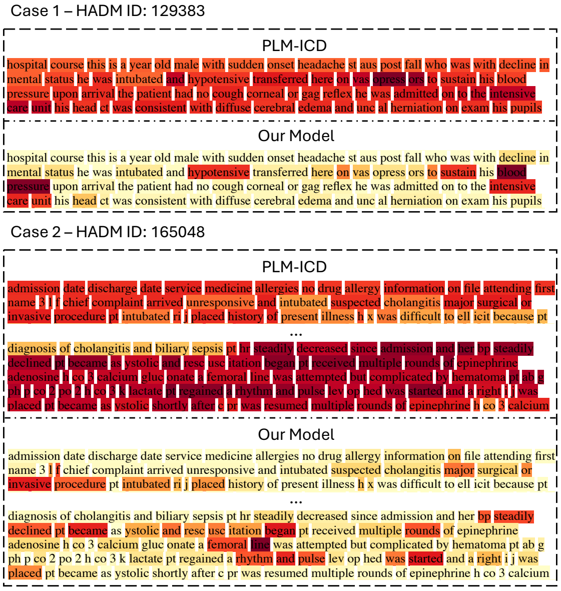


Figure 5: Visualisation of label-specific patients representation of codes 412 *'Old myocardial infarction'* and 39.95 *'Hemodialysis'*, without (left) and with (right) using knowledge graphs as input. Instances with the corresponding ground-truth label are red.



Case 1 – HADM ID: 129383

PLM-ICD

hospital course this is a year old male with sudden onset headache st aus post fall who was with decline in mental status he was intubated and hypotensive transferred here on vas opress ors to sustain his blood pressure upon arrival the patient had no cough corneal or gag reflex he was admitted on to the intensive care unit his head ct was consistent with diffuse cerebral edema and unc al herniation on exam his pupils

Our Model

hospital course this is a year old male with sudden onset headache st aus post fall who was with decline in mental status he was intubated and hypotensive transferred here on vas opress ors to sustain his blood pressure upon arrival the patient had no cough corneal or gag reflex he was admitted on to the intensive care unit his head ct was consistent with diffuse cerebral edema and unc al herniation on exam his pupils

Case 2 – HADM ID: 165048

PLM-ICD

admission date discharge date service medicine allergies no drug allergy information on file attending first name 3 l f chief complaint arrived unresponsive and intubated suspected cholangitis major surgical or invasive procedure pt intubated ri j placed history of present illness h x was difficult to ell icit because pt

...

diagnosis of cholangitis and biliary sepsis pt hr steadily decreased since admission and her bp steadily declined pt became as ystolic and resc usc itation began pt received multiple rounds of epinephrine adenosine h co 3 calcium gluc onate a femoral line was attempted but complicated by hematoma pt ab g ph p co 2 po 2 h co 3 k lactate pt regained a rhythm and pulse lev op hed was started and a right i j was placed pt became as ystolic shortly after c pr was resumed multiple rounds of epinephrine h co 3 calcium

Our Model

admission date discharge date service medicine allergies no drug allergy information on file attending first name 3 l f chief complaint arrived unresponsive and intubated suspected cholangitis major surgical or invasive procedure pt intubated ri j placed history of present illness h x was difficult to ell icit because pt

...

diagnosis of cholangitis and biliary sepsis pt hr steadily decreased since admission and her bp steadily declined pt became as ystolic and resc usc itation began pt received multiple rounds of epinephrine adenosine h co 3 calcium gluc onate a femoral line was attempted but complicated by hematoma pt ab g ph p co 2 po 2 h co 3 k lactate pt regained a rhythm and pulse lev op hed was started and a right i j was placed pt became as ystolic shortly after c pr was resumed multiple rounds of epinephrine h co 3 calcium

Figure 6: Highlights related to label 38.91 *'Arterial Catheterization'*, without (above) and with (below) using knowledge graphs as input.

with the target label (red), with an average distance of 16.44 between positive points compared to 19.14 for PLM-ICD. For code 39.95, where both models perform well, our model still shows a denser cluster of the positive (red) instances, and the cluster is more distinctly separated from other points. This case study demonstrates that integrating structured information enhances patient representation, leading to more accurate classification.

D. Explainability The ability to provide trustworthy and interpretable explanations is particularly critical in the clinical domain. To achieve this, we highlight text spans based on their attention weights, using darker colors to indicate higher weights. This suggests that these spans contribute more significantly to representing the patient. Our model demonstrates the ability to identify the most relevant spans more accurately and concisely. To illustrate this, we present two non-cherry-picked examples from the test set on label 38.91: *'Arterial Catheterization'*, where our model shows the most improvement. In Case 1 (Figure 6, above), our model effectively captures key tokens like *'hypotensive'* and *'blood pressure'*, which are directly associated with *'Arterial Catheterization'*, whose role is continuous blood pressure monitoring and arterial blood gas analysis. In contrast, PLM-ICD distributes attention more evenly across the text. In Case 2 (Figure 6, below), our model successfully highlights relevant spans across various sections, such as *'invasive procedure'* and *'placing a femoral line'*, they are procedures often involved in *'Arterial Catheterization'*. Additionally, phrases like *'intubated rij placed'* and *'a right IJ was placed'* are highlighted as they pertain to *'central venous catheterization'*, which is another type of catheterization. The model also succinctly highlights *'rhythm and pulse'*, which is related to blood pressure monitoring. These two cases strongly demonstrate that our model excels in providing high-quality explanations compared to PLM-ICD.

5 Conclusion

In this work, we construct a patient-level knowledge graph comprising wide range of entities and relationships. We integrate it into a state-of-the-art ICD coding architecture, PLM-ICD, which significantly enhances the patient representation and improve the coding performance. Additionally, we verify the impact of different types of entities and relationships in representing the patient. Furthermore, we showcase how integrating graph improves the patient representation through visualisation and demonstrate the high-quality explainability of our model in case studies.

Our patient-level knowledge graph dataset holds significant potential to provide healthcare providers with more precise, data-driven insights, ultimately improving patient outcomes, such as optimizing treatment plans and enabling early diagnosis.

6 Limitations

While our model demonstrates the effectiveness of integrating graph-based information, more advanced models have been developed in parallel (e.g., (Nguyen et al., 2023)). Our results confirm the benefits of structured knowledge integration by comparing our model to its baseline, highlighting the quality of our patient KG. However, we acknowledge that integrating structured information into these more advanced models could further enhance their performance, which remains an avenue for future exploration.

This work serves as the evaluation of the constructed patient-level KG from two perspectives: statistical and representational. However, it lacks a comprehensive comparison with other types of patient KGs (except for Yuan et al. (2021)) due to the challenges and time constraints associated with constructing them based on MIMIC-III.

In future work, we will also aim to enrich the patient-level knowledge graph by integrating other knowledge sources, such as hierarchical information from ontology systems like SNOMED-CT and UMLS. In the current study, we did not account for the semantic meaning of edges within graph representation, as some links merely signify connections between entities (e.g., *'1'* or *'TREATMENT-TEST'*). Moving forward, we plan to model the meaning of these relationships more explicitly by combining their semantic representations with confidence measurements.

Additionally, we have not explored other advanced graph representation models, such as Relational Graph Convolutional Networks (R-GCN) (Schlichtkrull et al., 2018) and Graph Attention Networks (GAT) (Veličković et al., 2017). The application of GAT, in particular, offers potential for further enhancing explainability by identifying and highlighting the sub-graphs that contribute most to final predictions, which we aim to evaluate more rigorously in domain expert-centred experiments.

Finally, due to resource constraints, we have not experimented with adapting other baseline models to use the document-level structured representation graphs. It is unlikely, but not impossible, that other architectures would not benefit from this kind of information, and further experiments should be conducted to establish this fact empirically.

References

James Blundell. 2023. Health information and the importance of clinical coding. *Anaesthesia & Intensive Care Medicine*.

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*, pages 3105–3114.

Finneas Catling, Georgios P Spithourakis, and Sebastian Riedel. 2018. Towards automated clinical coding. *International journal of medical informatics*, 120:50–61.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels. *arXiv preprint arXiv:2010.01653*.

Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Talukdar, and Steven Carroll. 2007. Automatic code assignment to medical text. In *Biological, translational, and clinical language processing*, pages 129–136.

Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728.

Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on mimic-iii and mimic-iv: a critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2572–2582.

Matús Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. 2022. Horses to zebras: ontology-guided data augmentation and synthesis for icd-9 coding. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics.

Matúš Falis, Maciej Pajak, Aneta Lisowska, Patrick Schrempf, Lucas Deckers, Shadia Mikhael, Sotirios Tsaftaris, and Alison O'Neil. 2019. Ontological attention ensembles for capturing semantic concepts in icd code prediction from clinical text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 168–177.

Malte Feucht, Zhiliang Wu, Sophia Althammer, and Volker Tresp. 2021. Description-based label attention classifier for explainable icd-9 classification. *arXiv preprint arXiv:2109.12026*.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. Plm-icd: automatic icd coding with pretrained language models. *arXiv preprint arXiv:2207.05289*.

Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the magic of bert apply to medical code assignment? a quantitative study. *Computers in biology and medicine*, 139:104998.

Shaoxiong Ji, Xiaobo Li, Wei Sun, Hang Dong, Ara Taalas, Yijia Zhang, Honghan Wu, Esa Pitkänen, and Pekka Marttinen. 2022. A unified review of deep learning for automated medical coding. *ACM Computing Surveys*.

John Snow Labs. 2024. Healthcare NLP. https://www.johnsnowlabs.com/healthcare-nlp/.

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

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd clinical natural language processing workshop*, pages 146–157.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8180–8187.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

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.

Jueqing Lu, Lan Du, Ming Liu, and Joanna Dipnall. 2020. Multi-label few/zero-shot learning with knowledge aggregated from multiple label graphs. *arXiv preprint arXiv:2010.07459*.

George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. 2022. Icdbigbird: a contextual embedding model for icd code classification. *arXiv preprint arXiv:2204.10408*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Anthony N Nguyen, Donna Truran, Madonna Kemp, Bevan Koopman, David Conlan, John O'Dwyer, Ming Zhang, Sarvnaz Karimi, Hamed Hassanzadeh, Michael J Lawley, et al. 2018. Computer-assisted diagnostic coding: effectiveness of an nlp-based approach using snomed ct to icd-10 mappings. In *AMIA Annual Symposium Proceedings*, volume 2018, page 807. American Medical Informatics Association.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Kashyap, and Stefan Winkler. 2023. A two-stage decoder for efficient icd coding. *arXiv preprint arXiv:2306.00005*.

World Health Organization et al. 1978. *International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index*. World Health Organization.

Suzanne Pereira, Aurélie Névéol, Philippe Massari, Michel Joubert, and Stefan Darmoni. 2006. Construction of a semi-automated icd-10 coding help system to optimize medical and economic coding. In *MIE*, pages 845–850. Citeseer.

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

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer.

Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric Xing. 2021. Generalized zero-shot text classification for icd coding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4018–4024.

Wei Sun, Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2021. Multitask recalibrated aggregation network for medical code prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 367–383. Springer.

Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. 2022. A review on deep neural networks for icd coding. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4357–4375.

Betty Van Aken, Jens-Michalis Papaioannou, Marcel G Naik, Georgios Eleftheriadis, Wolfgang Nejdl, Felix A Gers, and Alexander Löser. 2022. This patient looks like that patient: Prototypical networks for interpretable diagnosis prediction from clinical text. *arXiv preprint arXiv:2210.08500*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. *arXiv preprint arXiv:2007.06351*.

Tao Wang, Linhai Zhang, Chenchen Ye, Junxi Liu, and Deyu Zhou. 2022. A novel framework based on medical concept driven attention for explainable medical code prediction via external knowledge. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1407–1416.

Xiancheng Xie, Yun Xiong, Philip S Yu, and Yangyong Zhu. 2019. Ehr coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 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 *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2022, page 1767. NIH Public Access.

Quan Yuan, Jun Chen, Chao Lu, and Haifeng Huang. 2021. The graph-based mutual attentive network for automatic diagnosis. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3393–3399.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic icd coding. *arXiv preprint arXiv:2203.01515*.

Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. 2018. An end-to-end deep learning architecture for graph classification. In *AAAI*.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. Bert-xml: Large scale automated icd coding using bert pretraining. *arXiv preprint arXiv:2006.03685*.

A Appendix

A.1 Patient-Level Knowledge Graph Construction

Model Selection The Healthcare NLP library includes 44 RE models, each integrating both NER and RE functionalities. These models are trained on various language models across multiple languages to extract a wide range of clinical information. We utilize 14 of these models, which cover all available relationship types except for *'drug-drug interaction'* and share a consistent architecture. Details and statistics of these RE models are provided in Table 5.

The top five relationships include *ade conversational'*, which links drugs to their adverse reactions. However, we do not select it due to its uneven distribution across samples, as only a limited number contain this type of triple. Instead, we chose the *'bodypart-problem'* relationship, which ranks sixth.

Selected 5 RE Models Table 8 details the entities and relationships that each RE model can extract. The entities recognized from the MIMIC-III notes include a subset of those listed in this table.

Statistics of Entities Extracted The complete patient-level knowledge graph, which encompasses all five relationships, identifies 14 types of entities. The statistics of them can be found in Table 6. In the ablation study, we study the impact of top four types of entities, as they have the highest magnitude compared to others.

A.2 Patient-Level Knowledge Graph Visualisation

Figure 8 presents a visualization of a patient-level knowledge graph (HADM ID: 196292). To make it clear, we include type information for the entities, linking each entity to its respective type. Nodes representing types are colored light green, while different types of entities assigned unique colors.

A.3 Methodology of Information Entropy and Results of Ablation Study

Information Entropy Information entropy, introduced by Shannon in 1948, is a fundamental concept in information theory that measures information loss by quantifying the difference between the expected information and the reduced information. The entropy H of a discrete source X is given by:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x). \tag{10}$$

The entropy of text and serialised graph are calculated as follows:

$$H_{text} = -\sum_{x \in X_{text}} P_{text}(x) \log_2 P_{text}(x), \tag{11}$$

$$H_{graph} = -\sum_{x \in X_{graph}} P_{graph}(x) \log_2 P_{graph}(x). \tag{12}$$

The ratio of information loss L is defined as:

$$L = \frac{H_{text} - H_{graph}}{H_{text}}. \tag{13}$$

RE Model	$\|Tr\|$	$\|S\|$
clinical relatioship	6878467	52721
temporal events	6504349	52720
posology relationship	3939341	51879
ade conversational relationship	2443125	12464
bodypart-directions	355260	42487
bodypart-problem	337041	38719
ade relationship	86062	24259
test-problem-finding	76262	29007
drugprot relationship	42071	16859
bodypart-proceduretest	14739	8861
generic relationship	7004	2897
date relationship	2979	1713
test-result-date	2174	2174
phenotype gene relationship	0	0

Table 5: Statistics of RE model outputs in the MIMIC-III Full dataset. $|Tr|$ refers to the number of triples recognized by the RE model. $|S|$ indicates the number of samples in the full dataset that contain these triples.

Entity Type	$\|En\|$
problem	3422556
treatment	1665523
test	1371889
drug	1039115
strength	636491
frequency	338332
form	229420
dosage	217178
internal organ or component	192503
route	166454
direction	135903
symptom	106114
external body part or region	86367
duration	41727

Table 6: Statistics of entities identified in the MIMIC-III Full dataset. $|En|$ represents the number of entities.

Ablation Study Table 7 displays the information entropy results for different graphs after removing one type of relationship or entity. The *'Text Entropy'* is 8.33 across all experiments. Notably, the removing *'posology relationship'* and *'problem'* have the most significant impact on the results. This analysis emphasizes the loss of textual information, whereas the ablation study in the main content examines the impact on ICD coding.

A.4 Implementation Details and Results of Various DGCNN Configurations

Table 9 outlines the hyperparameter settings for both the PLM-ICD baseline and our model. Our model requires a batch size of 1 per process, as we do not adjust the graph representation using padding, unlike typical text inputs. Due to computational resource constraints, we do not use the optimal hyperparameters for PLM-ICD. However, we maintain consistent hyperparameters within their shared architecture to ensure a fair comparison. The value of DGCNN indicates the size of the node representation for each convolution layer. A single DGCNN layer with a size of 768 achieves the best performance on the full dataset, while two DGCNN layers, each with a size of 384, performs best on the Top-50 dataset. Additionally, we initialize the node representation in the first layer using RoBERTa-base.

Tables 10 and 11 present additional experimental

Remove	Graph Entropy	Ratio (%)
Full	7.48	89.95
clinical relationship	7.42	89.07
temporal events	7.33	88.07
posology relationship	7.15	85.80
bodypart-directions	7.47	89.68
bodypart-problem	7.48	89.80
problem	6.80	81.62
treatment	7.27	87.25
test	7.27	87.30
drug	7.36	88.40

Table 7: Results of the ablation study on information entropy: impact of removing each type of relationship or entity (MIMIC-III Full).

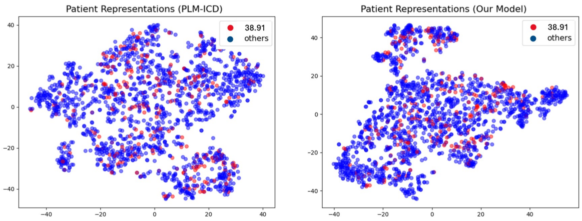


Figure 7: Visualisation of label-specific patients representation of code 38.91 *'Arterial Catheterization'*, without (left) and with (right) using knowledge graphs as input. Instances with the corresponding ground-truth label are red.

results for different configurations of the DGCNN architecture. The experiments utilize a complete graph with five types of relationships. In our initial experiment, we fix the final node size at 768 and compare the performance of DGCNN with different numbers of layers. The results indicate minimal performance differences between multi-layer and single-layer DGCNN models. However, models with evenly distributed layer sizes show slightly better performance. We also conduct experiments by varying the final node size and incrementally adding layers, each with an embedding size of 384. The results reveal an initial increase in performance, which subsequently decreases, with the optimal performance observed using two layers. Additionally, a similar trend is evident in a third experiment, which investigates varying sizes for each layer.

A.5 Patient visualisation - Code 38.91

We present a negative example of patient visualization for code 38.91 *'Arterial Catheterization'* in Figure 7, where both models exhibit poor performance. Our model achieves an F1-score of 38.14%, compared to 18.89% for PLM-ICD.

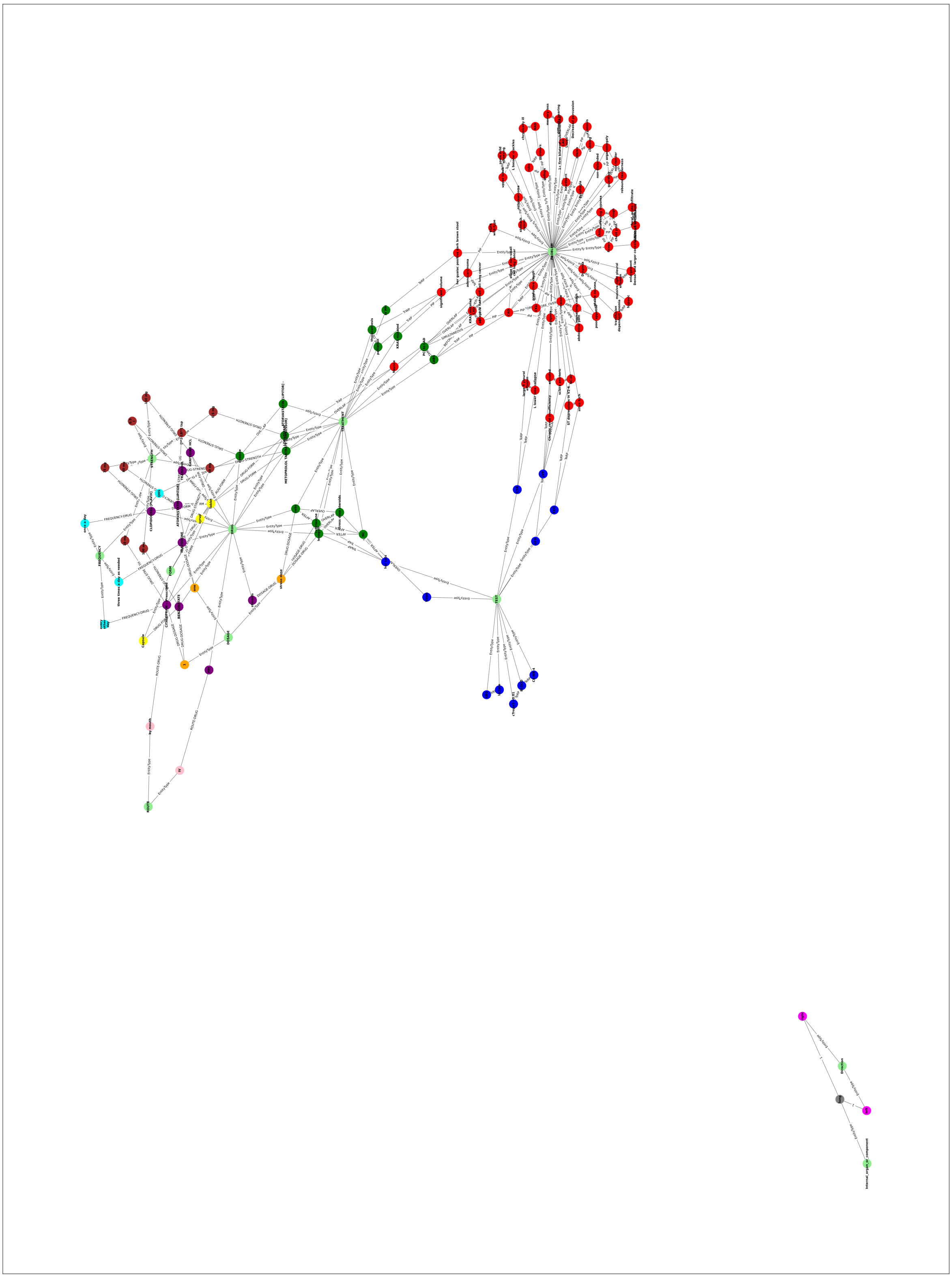


Figure 8: Visualisation of a Patient-Level Knowledge Graph.

RE Model	Entity	Relationship
clinical relationship	PROBLEM, TREATMENT, TEST	TrAP: TREATMENT-PROBLEM TeRP: TEST-PROBLEM TrIP: TREATMENT-PROBLEM TrCP: TREATMENT-PROBLEM TeCP: TEST-PROBLEM TrWP: TREATMENT-PROBLEM PIP: PROBLEM-PROBLEM O: No Relationship
temporal events	EVIDENTIAL, OCCURRENCE , DATE, TREATMENT, TIME, ADMISSION, TEST, FREQUENCY, CLINICAL_DEPT, DURATION, PROBLEM, DISCHARGE	BEFORE, AFTER, OVERLAP
posology relationship	drug, dosage, duration, strength, frequency	DOSAGE-DRUG DRUG-DURATION DRUG-STRENGTH DRUG-FREQUENCY
bodypart-directions	direction-external_body_part_or_region, external_body_part_or_region-direction, direction-internal_organ_or_component, internal_organ_or_component-direction	1,0
bodypart-problem	link between external_body_part_or_region or internal_organ_or_component and diseases entities (cerebrovascular_disease , communicable_disease, diabetes...)	1,0

Table 8: Entities and relationships that RE models can extract.

Input	Parameter	Value
Common	number of processes	4
	train/evaluation batch size	1
	gradient accumulation steps	1
	train epochs	20 (Full) / 10 (Top-50)
	warmup steps	2000
	random seed	42
Text	max length	5120
	chunk size	512
	model mode	LAAT
	pretrained model (text)	RoBERTa-base-PM
Graph	DGCNN	768 (Full) / 384-384 (Top-50)
	pretrained model (node)	RoBERTa-base

Table 9: Parameter settings for PLM-ICD (Common + Text) and our model (Common + Text + Graph) on the MIMIC-III Full and Top-50 datasets.

		F1		AUC		Precision	Recall
3-4 6-7 **Model**	**Embedding Size**	Macro	Micro	Macro	Micro	P@8	R@8
1 layer	768	11.05	59.72	92.37	98.75	76.59	40.52
	256-512	10.69	59.52	92.42	98.78	76.79	40.56
2 layers	384-384	10.98	59.64	92.65	98.83	76.63	40.53
	128-256-384	10.42	59.51	92.55	98.79	76.47	40.39
3 layers	256-256-256	10.53	59.70	92.47	98.84	76.81	40.56
	128-128-256-256	10.46	59.47	92.23	98.78	76.58	40.37
4 layers	192-192-192-192	10.58	59.21	92.14	98.78	76.47	40.37
1 layer	384	10.77	59.77	92.30	98.77	76.88	40.62
2 layers	384-384	10.82	59.43	92.54	98.78	76.63	40.53
3 layers	384-384-384	10.49	59.37	92.35	98.75	76.11	40.15
4 layers	384-384-384-384	10.46	59.58	92.23	98.74	76.79	40.55
1 layer	128	10.23	59.06	92.22	98.82	76.65	40.46
2 layers	128-256	10.60	59.69	92.47	98.83	76.85	40.57
3 layers	128-256-384	10.42	59.51	92.55	98.79	76.47	40.39
4 layers	128-256-384-512	10.47	59.31	92.21	98.76	76.30	40.27

Table 10: Performance of Various DGCNN Architecture Configurations (MIMIC-III Full).

		F1		AUC		Precision	Recall
3-4 6-7 **Model**	**Embedding Size**	Macro	Micro	Macro	Micro	P@5	R@5
1 layer	768	66.64	71.37	91.77	94.16	66.52	64.33
	256-512	67.64	71.72	92.12	94.30	66.82	64.74
2 layers	384-384	67.81	71.63	92.04	94.22	67.08	65.11
	128-256-384	66.30	70.94	91.71	93.98	66.58	64.39
3 layers	256-256-256	67.54	71.83	92.19	94.37	67.04	65.14
	128-128-256-256	66.79	71.39	92.28	94.31	66.87	64.92
4 layers	192-192-192-192	67.67	72.03	92.30	94.44	66.79	65.07
1 layer	384	66.91	71.12	92.04	94.19	66.47	64.63
2 layers	384-384	67.81	71.63	92.04	94.22	67.08	65.11
3 layers	384-384-384	66.89	71.41	92.26	94.32	67.09	65.20
4 layers	384-384-384-384	66.55	71.24	92.15	94.37	66.86	64.85
1 layer	128	66.62	70.79	91.89	94.10	66.50	64.45
2 layers	128-256	67.63	71.72	92.00	94.25	66.71	64.63
3 layers	128-256-384	66.30	70.94	91.71	93.98	66.58	64.39
4 layers	128-256-384-512	65.80	71.34	92.05	94.32	66.14	64.22

Table 11: Performance of Various DGCNN Architecture Configurations (MIMIC-III Top-50).