Automatic ICD Coding with Pretrained Language Models

Shijie Zhou[1], **Fan Zhang**[1], **Yilong Chen**[2], **Bosheng Chen**[2], **Liping Huang**[2]

[1]School of Information, Xiamen University
[2]Artificial Intelligence Research Institute, Xiamen University
{31520241154542, 31520241154537, 36920241153201, 36920241153196, 36920241153215}@stu.xmu.edu.cn

Abstract

Accurately assigning medical codes from International Classification of Diseases(ICD) to free-text medical documentation is crucial for healthcare operations but traditionally requires significant manual effort. Automatic ICD coding using NLP is gaining interest, particularly with pretrained language model (PLM). This research enhances PLM-based ICD coding, inspired by the PLM-ICD framework and the Read, Attend, and Code (RAC) model. This work focuses on addressing the challenge of the large label space in ICD coding. A novel framework is proposed that integrates the code-title guided attention modules from RAC into the PLM-ICD architecture. This approach leverages code title information to learn semantic relationships between clinical notes and ICD codes. Experiments conducted on the MIMIC-III dataset demonstrate the effectiveness of proposed approach. The modified framework demonstrates superior performance compared to the original PLM-ICD. The integration of domain-specific knowledge and advanced attention mechanisms improves the accuracy of Automatic ICD coding.

Introduction

Automatic ICD coding coding is the task of assigning diagnosis and procedure codes to free-text medical documentation (Dong et al. 2022). These codes ensure that patients receive the correct level of care and that healthcare providers are accurately compensated for their services. However, this is a costly manual process prone to error (O'Malley et al. 2005; Tseng et al. 2018). The goal of automatic ICD coding is to predict a set of codes or provide a list of codes ranked by relevance for a medical document. Numerous machine learning models have been developed for automatic ICD coding (Stanfill et al. 2010) . These models are trained on datasets of medical documents, typically discharge summaries, each labeled with a set of medical codes. While some models treat automatic ICD coding as an ad-hoc information retrieval problem (Rizzo et al. 2015; Park et al. 2019), it is more commonly posed as a multi-label classification problem.

Prior work has identified several challenges of this task, including the large number of labels to be classified, the long input sequence, and the imbalanced label distribution, i.e., the long-tail problem (Xie et al. 2019). Taking automatic international classification of diseases (ICD) coding as example, given discharge summaries notes as input, the task is to assign multiple ICD disease and procedure label codes associated with each note. The assigned codes need to be accurate and complete for the billing purposes. To mitigate the data sparsity problem, additional structured knowledge could be applied. The textual description of medical codes describes the exact meaning of codes and provides extra semantic information for abstract codes.

Recently, pretrained language models (PLMs) with the Transformer (Vaswani et al. 2017)architecture have become the dominant forces for NLP research. These models are pretrained on large amount of text with various language modeling objectives, and then fine-tuned on the desired downstream tasks to perform different functionalities.

Thus, more researchers have proposed to use transformer-based models. Zhang, Liu, and Razavian proposed BERT-XML that combines BERT encoders with multi-label attention. Huang et al.Huang, Tsai, and Chen developed a Transformer-based pretrained language model with domain-specific PLM and segment pooling for the long input sequence problem. Kim and Ganapathi implements the code-title guided attention module.Yang et al. adopted longformer with domain-specific knowledge enhancement.

In this paper, we propose a novel framework that integrates the code-title guided attention modules from RAC into the PLM-ICD architecture. Our main contributions are as follows:

- We propose a framework that combines the advantages of PLM-ICD and RAC models, specifically integrating code-title guided attention mechanisms into the PLM-based architecture for medical coding.
- We address the large label space challenge in medical coding by leveraging code title information to establish stronger semantic connections between clinical notes and medical codes.
- We conduct comprehensive experiments on the MIMIC-III dataset, demonstrating that our proposed framework achieves superior performance compared to the original PLM-ICD model, particularly in handling complex medical coding scenarios.

Related Work

Automated Medical Coding

ICD code prediction is a challenging task in the medical domain. Several recent work attempted to approach this task with neural models. Choi et al.; Baumel et al. used recurrent neural networks (RNN) to encode the EHR data for predicting diagnostic results. Li and Yu recently utilized a multi-filter convolutional layer and a residual layer to improve the performance of ICD prediction. On the other hand, several work tried to integrate external medical knowledge into this task. In order to leverage the information of definition of each ICD code, RNN and CNN were adopted to encode the diagnostic descriptions of ICD codes for better prediction via attention mech- anism (Shi et al. 2017; Mullenbach et al. 2018). Moreover, the prior work tried to consider the hi- erarchical structure of ICD codes (Xie and Xing 2018), which proposed a tree-of-sequences LSTM to simultaneously capture the hierarchical relation- ship among codes and the semantics of each code. Also,Tsai, Chang, and Chen introduced various ways of leveraging the hierarchical knowledge of ICD by adding refined loss functions. Recently, Cao et al. proposed to train ICD code embeddings in hyperbolic space to model the hierarchical struc- ture. Additionally, they used graph neural network to capture the code co-occurrences. LAAT (Vu, Nguyen, and Nguyen 2020) integrated a bidirectional LSTM with an improved label-aware attention mechanism. Ef- fectiveCAN (Liu et al. 2021) integrated a squeeze- and-excitation network and residual connections along with extracting representations from all encoder layers for label attention. The authors also introduced focal loss to tackle the long-tail predic- tion problem. ISD (Zhou et al. 2021)employed extraction of shared representations among high- frequency and low-frequency codes and a self- distillation learning mechanism to alleviate the long-tail code distribution. Kim and Ganapathi proposed a framework called Read, Attend, and Code (RAC) to effectively predict ICD codes, which is the current state-of-the-art model on this task. Most recent models focused on developing an effective interaction between note representations and code representations (Cao et al. 2020; Zhou et al. 2021; Kim and Ganapathi 2021).

Model architectures

Most recent state-of-the-art models use an encoder-decoder architecture. The encoder takes a sequence of tokens $T \in Z^n$ as input and outputs a sequence of hidden representations $H \in R_h^d \times n$, where n is the number of tokens in a sequence, and d_h is the hidden dimension. The decoder takes H as input and outputs the code probability distributions. For the task of ranking, codes are sorted by decreasing probability. For classification, code probabilities larger than a set decision boundary are predicted.

Encoders The encoder usually consists of pre-trained non-contextualized word embeddings (e.g., Word2Vec) and a neural network for encoding context. More recently, pre-trained masked language models (e.g., BERT) have gained popularity (Teng et al. 2023). In order to mitigate the domain mismatch problem, we propose to utilize the PLMs that are pretrained on biomedical and clinical text, e.g..The MIMIC-III training set or PubMed articles are commonly used for pre-training.

Decoders The most common decoder architectures can be grouped into three primary types. The simplest decoder is a pooling layer (e.g., max pooling) followed by a feed-forward neural network. More recently, label-wise attention (LA) (Mullenbach et al. 2018) has replaced pooling (Huang, Tsai, and Chen 2022; Li and Yu 2020; Liu et al. 2021; Vu, Nguyen, and Nguyen 2020). LA transforms a sequence of hidden representations H into label-specific representations $V \in R^{d_h \times L}$, where L is the number of unique medical codes in the dataset. It is computed as

$$A = \text{softmax}(WH), \quad V = HA^\top, \tag{1}$$

where the softmax normalizes each column of WH, $W \in R^{L \times d_h}$ is an embedding matrix that learns label-specific queries, and $A \in R^{L \times n}$ is the attention matrix. Then, V is used to compute class-wise probabilities via a feedforward neural network. As LA was first used in the convolutional attention for multi-label classification (CAML) model (Mullenbach et al. 2018), we refer to this method as LA_{CAML}.

An updated label-wise attention module was introduced in the label attention model (LAAT) (Vu, Nguyen, and Nguyen 2020). We refer to this attention module as LA_{LAAT}. In LA_{LAAT}, the label-specific attention is computed similarly to LA_{CAML} as $A = \text{softmax}(UZ)$, where $U \in R^{L \times d_p}$ is a learnable embedding matrix, but with $Z = \tanh(PH)$ where $P \in R^{d_p \times d_h}$ is a learnable matrix, $Z \in R^{d_p \times n}$ and d_p is a hyperparameter.

Usage of Auxiliary Information Auxiliary information can be utilized to enhance representation learning and improve the performance of medical coding. This section introduces the usage of auxiliary information, including implicit information such as label information via randomly initialized embeddings and explicit information (or external data) such as Wikipedia articles, textual code descriptions, and code hierarchies. Implicit label information has been used by most previously introduced label attention–based models. The joint embedding model (LEAM) (Wang et al. 2018) embeds labels and leverages the compatibility between word and label embeddings to calculate attention scores. The following paragraphs review the methods that use external data explicitly. The external data can be applied to both encoders and decoders. When applied to encoders, external data enhance the representation learning of clinical texts. The external information usually acts as the regularization for decoders when combining external data augmentation with the decoding process.

The textual description of medical codes describes the exact meaning of codes and provides extra semantic information for abstract codes. The embeddings of code description are denoted as , where m is the number of codes, and is the dimension of description embedding. Several publications utilize the code description to enhance representation learning. DR-CAML(Mullenbach et al. 2018) uses the word vectors of description as a regularization when optimizing the labelwise attention module. Similarly, CAIC (Teng et al.

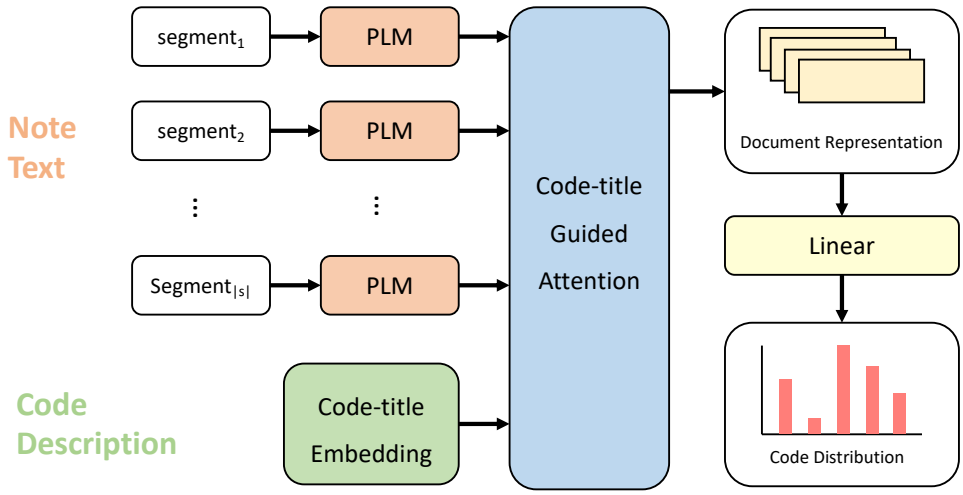

Figure 1: llustration of our proposed framework. LM encodes segments of a document separately, and a code-title guided attention mechanism is to aggregate the segment representations into label-aware document representations. The document representations are linear-transformed to predict ICD codes.

2020) develops cross-textual attention to establish the connection between medical notes and ICD codes. A prompt-based fine-tuning model (Yang et al. 2022) adds a series of ICD code descriptions as the prompt to integrate code description and input notes for multi-label few-shot ICD coding.

Proposed Solution

Our proposed framework combines the strengths of PLM-ICD and RAC, integrating three key components:

- Domain-specific PLM for encoding medical text
- Segment pooling to handle long documents
- Code-title guided attention for label-aware document representation

As shown in Figure 1, the framework first uses a domain-specific PLM to encode document segments. The segment pooling mechanism then aggregates these representations. Finally, code-title guided attention generates label-aware document representations for ICD code prediction.

Domain-Specific Pretraining Automatic ICD coding is a domain-specific task where the input text consists of clinical notes writ- ten by clinicians. The clinical notes contain many biomedical terms, and understanding these terms is essential in order to assign ICD codes accurately. While general PLMs are pretrained on large amount of text, the pretraining corpora usually does not contain biomedical text, not to mention clinical records. In order to mitigate the domain mismatch prob- lem, we propose to utilize the PLMs that are pretrained on biomedical and clinical text. These PLMs are specifically pretrained for biomedical tasks and proven to be effective on various downstream tasks. We take the domain-specific PLMs and fine-tune them on the task of automatic ICD coding. We can plug-and-play the domain- specific PLMs since their architectural design and pretraining objective are identical to their general- domain counterparts. This makes our framework agnostic to the type of PLMs, i.e., we can apply any transformer-based PLMs as the encoder.And RoBERTa-PM (Lewis et al. 2020) has the best performance in PLM-ICD model (Huang, Tsai, and Chen 2022).

Segment Pooling In order to tackle the long input text problem, Huang, Tsai, and Chen propose segment pooling to surpass the maximum length limitation of PLMs. The segment pooling mechanism first splits the whole document into segments that are shorter than the maximum length, and encodes them into segment representations with PLMs. After encod- ing segments, the segment representations are aggregated as the representations for the full docu- ment.

More formally, given a document $d = \{t_1, t_2, \ldots, t_{|d|}\}$ of $|d|$ tokens, we split it into $|s|$ consecutive segments s_i of c tokens:

$$s_i = \{t_j \mid c \cdot i \leq j < c \cdot (i+1)\}$$

The segments are fed into PLMs separately to compute hidden representations, then concatenated to obtain the hidden representations of all tokens:

$$H = \text{concat}(PLM(s_1), \cdots, PLM(s_{|s|}))$$

The token-wise hidden representations H can then be used to make predictions based on the whole document.

Code-Title Guided Attention (Kim and Ganapathi 2021) use the definition tables of the diagnoses and procedure codes, concatenate long and short titles together for all n_y codes, and build $\mathcal{C}_T$ first. By tokenizing $\mathcal{C}_T$ with n_t tokens, we have a title matrix T where $T \in R^{n_y \times n_t}$. From T input, the module extracts a code-title embedding of dimension d by using an embedding layer followed by a single CNN layer

Table 1: Comparison of MIMIC-III full and MIMIC-III 50 datasets.

	MIMIC-III full	**MIMIC-III 50**
Number of documents	52,723	11,368
Number of patients	41,126	10,356
Number of unique codes	8,929	50
Codes pr. instance: Median (IQR)	14 (10-20)	5 (3-8)
Words pr. document: Median (IQR)	1,375 (965-1,900)	1,478 (1,065-1,992)
Documents: Train/val/test [%]	90.5/3.1/6.4	71.0/13.8/15.2
Missing codes: Train/val/test [%]	2.7/66.4/54.3	0.0/0.0/0.0

Table 2: Results on the MIMIC-III 50test set (%).

Model	**AUC**		**F1**		**P@5**
	Macro	**Micro**	**Macro**	**Micro**	
MultiResCNN	89.7	93.8	62.2	67.3	63.4
LAAT	90.5	92.8	60.8	66.8	64.0
PLM-ICD	91.7	93.8	66.3	70.5	65.7
Ours	92.09	94.10	64.56	71.10	66.45

and Global Max Pooling layer. We let $E_t \in R^{n_y \times d}$ be the extracted code-title embedding matrix. In the model, each concatenated code title is padded to $n_t = 36$ tokens, the same pre-trained Word2Vec Skip-gram model weights that the reader used are loaded to initialize the embedding layer, and a single CNN layer with d filters, kernel size 10, and tanh activation function are used. However, we use PLM to get initialized embedding layer.

This function computes code-level attention over the reader output to attend to different parts for each code. We explicitly use E_t as a query matrix to guide where to attend from the reader output. Specifically, the approach leads to the following attention mechanism:

$$V_x = \text{Softmax}\left(\frac{E_t U_x^T}{\sqrt{d}}\right) U_x,$$

where $U_x = \text{SAM}(E_x)$ and $V_x \in R^{n_y \times d}$.

With attended V_x, finally, the module produces a code likelihood vector y as

$$y = \sigma(V_x W_3),$$

where $W_3 \in R^{d \times 1}$ and σ is the sigmoid function.

Experiments

Datasets

The Medical Information Mart for Intensive Care III (MIMIC-3) (Johnson et al. 2016) dataset is a benchmark dataset which contains text and structured records from a hospital ICU. We use the same setting as Mullenbach et al., where 47,724 dis- charge summaries are used for training, with 1,632 summaries and 3,372 summaries for val- idation and testing, respectively. There are 8,922 labels in the dataset.MIMIC-III full and 50 are commonly used splits. Table 1 shows the details of the two splits.MIMIC-III full contains all ICD-9 codes, while 50 only contains the top 50 most frequent codes (Shi et al. 2017; Mullenbach et al. 2018).

Implementation Details

We take the pretrained weights released by original authors without any modification. For the best PLM-ICD model, Huang, Tsai, and Chen use RoBERTa-base-PM-M3- Voc released by Lewis et al. During fine- tuning, we train our models for 20 epochs. AdamW is chosen as the optimizer with a learning rate of 5e 5. We employ a linear warmup schedule with 2000 warmup steps, and after that the learning rate decays linearly to 0 throughout training. The batch size is set to 16. All models are trained on a GTX 4090 GPU. We truncate discharge summaries to 3072 tokens due to memory consideration, and the length of each segment c is set to 128.

We evaluate our methods with commonly used metrics to be directly comparable to previous work. The metrics used are macro F1, micro F1, macro AUC, micro AUC, and precision@K, where K = 8, 15 for MIMIC-III full and K = 5 for MIMIC-III 50.

Baselines

MultiResCNN (Li and Yu 2020) encode free text with Multi-Filter Residual CNN, and applied label code attention mechanism to enable each ICD code to attend different parts of the document.

LAAT (Vu, Nguyen, and Nguyen 2020) applies the structured self-attention that projected the hidden representation via a linear transformation and non-linear activation.

PLM-ICD (Huang, Tsai, and Chen 2022) uses domain-specific pre-training models with segment pooling for the long input sequence problem.

Results

Table 2 shows the results on MIMIC-III 50. Our proposed model achieves the best performance across most metrics.

Specifically, we improve the macro-AUC by 0.39%, micro-AUC by 0.3%, micro-F1 by 0.6%, and P@5 by 0.75% compared to PLM-ICD. The improvements demonstrate that integrating code-title guided attention mechanisms with domain-specific PLM enhances the model's ability to capture relevant information from clinical notes.

The effectiveness of our approach can be attributed to two key factors: (1) the domain-specific pretraining helps the model better understand medical terminology and context, and (2) the code-title guided attention mechanism enables more precise connections between clinical notes and ICD codes. This is particularly evident in the improved macro metrics, suggesting better handling of less frequent codes.

Ablation Study

Table 3 provides analysis on factors that affect PLM's performance on automatic ICD coding.

Table 3: Ablation Study on the MIMIC-III 50 test set (%).

Model	Macro-F	Micro-F
Ours	89.7	93.8
(a) - domain pretraining	88.4	92.3
(b) - code-title guided attention	87.1	90.0

Effect of Pretrained Models

Huang, Tsai, and Chen have already test 4 models accrding to the BLURB(the Biomedical Language Understanding and Reasoning Benchmark)(Gu et al. 2021).And we select (Yasunaga et al., 2022) and BioClinicalBERT(Alsentzer et al., 2019) from BLURB to compare the performance of Pretrained Models.Table 4 shows RoBERTa-PM has best performance.

Table 4: Results with different PLMs on the MIMIC-III 50 test set (%).

Model	Macro-F	Micro-F
RoBERTa-PM	92.09	94.10
BioLinkBERT	89.99	92.85
BioClinicalBERT	88.90	91.78

Conclusion

In this paper, we introduced a novel framework for automatic ICD coding that combines the strengths of PLM-ICD and RAC models. By integrating code-title guided attention mechanisms with domain-specific pretrained language models, our approach achieves superior performance on the MIMIC-III dataset. The experimental results demonstrate the effectiveness of leveraging code title information to establish semantic connections between clinical notes and ICD codes. Future work could explore additional domain-specific knowledge integration and attention mechanisms to further improve performance.

References

Baumel, T.; Nassour-Kassis, J.; Cohen, R.; and Elhadad, M. 2017. Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment.

Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Liu, S.; and Chong, W. 2020. HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3105–3114. Online: Association for Computational Linguistics.

Choi, E.; Bahadori, M. T.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, 301–318. PMLR.

Dong, H.; Falis, M.; Whiteley, W.; Alex, B.; Matterson, J.; Ji, S.; Chen, J.; and Wu, H. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1): 159.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1): 2:1–2:23.

Huang, C.-W.; Tsai, S.-C.; and Chen, Y.-N. 2022. PLM-ICD: Automatic ICD Coding with Pretrained Language Models. In Naumann, T.; Bethard, S.; Roberts, K.; and Rumshisky, A., eds., *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 10–20. Seattle, WA: Association for Computational Linguistics.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1): 160035.

Kim, B.-H.; and Ganapathi, V. 2021. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, 196–208. PMLR.

Lewis, P.; Ott, M.; Du, J.; and Stoyanov, V. 2020. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In Rumshisky, A.; Roberts, K.; Bethard, S.; and Naumann, T., eds., *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 146–157. Online: Association for Computational Linguistics.

Li, F.; and Yu, H. 2020. ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 8180–8187. AAAI Press.

Liu, Y.; Cheng, H.; Klopfer, R.; Gormley, M. R.; and Schaaf, T. 2021. Effective Convolutional Attention Network for

Multi-label Clinical Document Classification. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5941–5953. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; and Eisenstein, J. 2018. Explainable Prediction of Medical Codes from Clinical Text. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1101–1111. New Orleans, Louisiana: Association for Computational Linguistics.

O'Malley, K. J.; Cook, K. F.; Price, M. D.; Wildes, K. R.; Hurdle, J. F.; and Ashton, C. M. 2005. Measuring Diagnoses: ICD Code Accuracy. *Health Services Research*, 40(5p2): 1620–1639.

Park, H.; Castañ O, J.; Á Vila, P.; Pé Rez, D.; Berinsky, H.; n; Gambarte, L.; Luna, D.; and Otero, C. 2019. An Information Retrieval Approach to ICD-10 Classification. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, 1564–1565. IOS Press.

Rizzo, S. G.; Montesi, D.; Fabbri, A.; and Marchesini, G. 2015. ICD Code Retrieval: Novel Approach for Assisted Disease Classification. In Ashish, N.; and Ambite, J.-L., eds., *Data Integration in the Life Sciences*, 147–161. Cham: Springer International Publishing. ISBN 978-3-319-21843-4.

Shi, H.; Xie, P.; Hu, Z.; Zhang, M.; and Xing, E. P. 2017. Towards Automated ICD Coding Using Deep Learning. Publication Title: ArXiv preprint Volume: abs/1711.04075.

Stanfill, M. H.; Williams, M.; Fenton, S. H.; Jenders, R. A.; and Hersh, W. R. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6): 646–651.

Teng, F.; Liu, Y.; Li, T.; Zhang, Y.; Li, S.; and Zhao, Y. 2023. A Review on Deep Neural Networks for ICD Coding. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4357–4375.

Teng, F.; Yang, W.; Chen, L.; Huang, L.; and Xu, Q. 2020. Explainable Prediction of Medical Codes With Knowledge Graphs. *Frontiers in Bioengineering and Biotechnology*, 8.

Tsai, S.-C.; Chang, T.-Y.; and Chen, Y.-N. 2019. Leveraging Hierarchical Category Knowledge for Data-Imbalanced Multi-Label Diagnostic Text Understanding. In Holderness, E.; Jimeno Yepes, A.; Lavelli, A.; Minard, A.-L.; Pustejovsky, J.; and Rinaldi, F., eds., *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 39–43. Hong Kong: Association for Computational Linguistics.

Tseng, P.; Kaplan, R. S.; Richman, B. D.; Shah, M. A.; and Schulman, K. A. 2018. Administrative Costs Associated With Physician Billing and Insurance-Related Activities at an Academic Health Care System. *JAMA*, 319(7): 691–697.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. v.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Vu, T.; Nguyen, D. Q.; and Nguyen, A. 2020. A Label Attention Model for ICD Coding from Clinical Text. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 3335–3341. ijcai.org.

Wang, G.; Li, C.; Wang, W.; Zhang, Y.; Shen, D.; Zhang, X.; Henao, R.; and Carin, L. 2018. Joint embedding of words and labels for text classification. *arXiv preprint arXiv:1805.04174*.

Xie, P.; and Xing, E. 2018. A Neural Architecture for Automated ICD Coding. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1066–1076. Melbourne, Australia: Association for Computational Linguistics.

Xie, X.; Xiong, Y.; Yu, P. S.; and Zhu, Y. 2019. EHR Coding with Multi-scale Feature Attention and Structured Knowledge Graph Propagation. In Zhu, W.; Tao, D.; Cheng, X.; Cui, P.; Rundensteiner, E. A.; Carmel, D.; He, Q.; and Yu, J. X., eds., *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, 649–658. ACM.

Yang, Z.; Wang, S.; Rawat, B. P. S.; Mitra, A.; and Yu, H. 2022. Knowledge Injected Prompt Based Fine-tuning for Multi-label Few-shot ICD Coding. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 1767–1781. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhang, Z.; Liu, J.; and Razavian, N. 2020. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. In Rumshisky, A.; Roberts, K.; Bethard, S.; and Naumann, T., eds., *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 24–34. Online: Association for Computational Linguistics.

Zhou, T.; Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Niu, K.; Chong, W.; and Liu, S. 2021. Automatic ICD Coding via Interactive Shared Representation Networks with Self-distillation Mechanism. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5948–5957. Online: Association for Computational Linguistics.