Bridging the Code Gap: A Joint Learning Framework Across Medical Coding Systems

Geunyeong Jeong[1], Seokwon Jeong[2], Juoh Sun[1] Harksoo Kim[1,*]
[1]Konkuk University, Republic of Korea
[2]Kangwon National University, Republic of Korea
jyjg7218@konkuk.ac.kr, nlpsw@kangwon.ac.kr
qssz1326@konkuk.ac.kr, nlpdrkim@konkuk.ac.kr

Abstract

Automated Medical Coding (AMC) is the task of automatically converting free-text medical documents into predefined codes according to a specific medical coding system. Although deep learning has significantly advanced AMC, the class imbalance problem remains a significant challenge. To address this issue, most existing methods consider only a single coding system and disregard the potential benefits of reflecting the relevance between different coding systems. To bridge this gap, we introduce a Joint learning framework for Across Medical coding Systems (JAMS), which jointly learns different coding systems through multi-task learning. It learns various representations using a shared encoder and explicitly captures the relationships across these coding systems using the medical code attention network, a modification of the graph attention network. In the experiments on the MIMIC-IV ICD-9 and MIMIC-IV ICD-10 datasets, connected through General Equivalence Mappings, JAMS improved the performance consistently regardless of the backbone models. This result demonstrates its model-agnostic characteristic, which is not constrained by specific model structures. Notably, JAMS significantly improved the performance of low-frequency codes. Our analysis shows that these performance gains are due to the connections between the codes of the different coding systems.

Keywords: Automated Medical Coding, Medical Coding System, Joint Learning Framework

1. Introduction

Medical coding is the task of converting free-text medical documents into predefined codes according to a specific coding system, such as the International Classification of Diseases (ICD). This process standardizes the medical information to improve its accuracy and consistency, and it is used in various medical services (Choi et al., 2016) and insurance claims (Park et al., 2000). However, training human coders for medical coding is expensive. For example, training coders for national health services worldwide takes several months (Varela et al., 2022). Additionally, manual coding by human coders is time-consuming (Park et al., 2000) and prone to human error (O'Malley et al., 2005). Therefore, researchers have begun to automate medical coding processes. Automated Medical Coding (AMC) has significantly advanced through deep learning (Yuan et al., 2022; Yang et al., 2023). However, several challenges remain to be addressed. The class imbalance problem is a significant issue in AMC. The datasets used for AMC (Johnson et al., 2016; Jeong et al., 2023) are collected from real-world medical environments. Codes for common diseases, such as respiratory infections and coughs, appear frequently, while codes for rare diseases are scarce (Yan et al., 2022). This phenomenon leads to a long-tailed distribution of the codes. Using imbalanced datasets without appropriate strategies can bias the model toward high-frequency codes. To address this issue, researchers have proposed leveraging the hierarchical structure of the coding system (Vu et al., 2020; Nguyen et al., 2023) or enhancing label embedding methods for low-frequency codes (Zhang et al., 2022). However, most of these approaches focused on a single coding system.

We believe reflecting the relevance of different coding systems can alleviate the class imbalance problem. For instance, the ICD-9 code 405.11 (Benign renovascular hypertension) and the ICD-10 code I15.0 (Renovascular hypertension) are closely related. We expect the model to be effectively generalized by capturing the relationships between these codes. Therefore, we propose a Joint learning framework Across Medical coding Systems (JAMS)[1] that jointly learns different coding systems through multi-task learning, incorporating explicit mapping information across two systems. In this study, we used two medical coding systems, ICD-9 and ICD-10, and employed General Equivalence Mappings (GEMs)[2] to connect them. GEMs are official mapping tools between the ICD-9 and ICD-10 developed by various medical organizations, including the National Center for Health Statistics. The main contributions of our study are as follows:

* Corresponding author

[1] Our code is publicly available at `https://github.com/GY-Jeong/JAMS`.

[2] `https://www.cms.gov/medicare/coding-billing/icd-10-codes/2018-icd-10-cm-gem`

LREC-COLING 2024, pages 2520–2525
20-25 May, 2024.

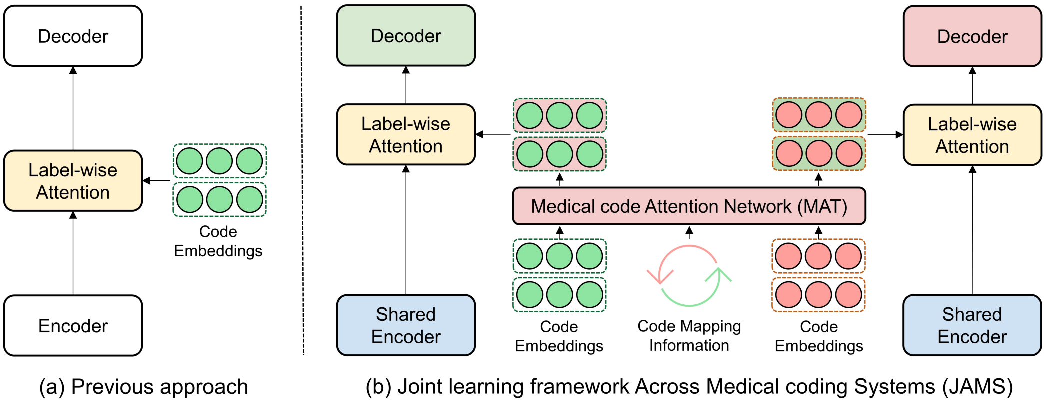

Figure 1: An illustration of (a) previous approach and (b) proposed approach (JAMS).

- While some studies have applied multi-task learning to the medical field (Suk et al., 2016; Sun et al., 2022), few have focused on AMC. Our study is one of the few multi-task learning methods available for AMC.

- JAMS is a model-agnostic approach that can be applied to any AMC model that adopts label-wise attention.

- JAMS significantly improves the predictions for low-frequency codes. In the medical field, accurately predicting low-frequency codes (e.g., rare diseases) is particularly important, making this a key contribution.

2. Methodology

Most of the recently proposed AMC models represent each code as a vector and perform label-wise attention with the encoder output to predict the codes (Zhang et al., 2022). Figure 1 (a) illustrates this approach. These models only consider a single coding system. In contrast, JAMS extends these models to learn from datasets with different coding systems jointly. Figure 1 (b) illustrates JAMS. JAMS employs a shared encoder to learn various representations from two datasets and adopts a graph network to leverage explicit mapping information across the two systems. This strategy allows JAMS to learn and represent interactions between medical coding systems more effectively.

2.1. Enhancing Code Embedding with Medical Code Attention Network

The GEMs are code mapping information designed to bridge the gap between the ICD-9 and the ICD-10. To leverage this information effectively, we propose the Medical Code Attention Network (MAT), a modification of the Graph Attention Network (GAT) introduced by Veličković et al. (2018). The GAT employs an attention mechanism to dynamically highlight and assess the importance of interactions between nodes in a graph. In our approach, we conceptualize each medical code as a node and leverage the code connection information from GEMs as edges. To make GAT more suitable, we modified it to capture the relationships between codes more accurately. There are two distinctions between MAT and GAT. First, while the standard GAT uses a single label-embedding table, MAT employs separate label-embedding tables for each coding system. This strategy allows the model to reflect the unique characteristics of each coding system more accurately. Second, MAT incorporates the "approximate flag" information from GEMs. This flag indicates whether two codes are "completely equivalent" or "approximately equivalent", providing a detailed understanding of their relationship. For instance, ICD-10 code K83.1 (Obstruction of bile duct) and ICD-9 code 576.2 (Obstruction of bile duct) are "completely equivalent" as both codes share the same meaning. Conversely, ICD-9 code 749.11 (Cleft lip, unilateral, complete) and ICD-10 code Q36.9 (Cleft lip, unilateral) are "approximately equivalent" because they are similar but not identical. By integrating this, we can model the nuanced relationship between codes across the two systems more precisely.

The specific method is as follows: we define the embedding table for ICD-9 codes as $U = \{u_i\}_{i=1}^{n} \in \mathbb{R}^{n\times d}$, and for ICD-10 codes as $V = \{v_j\}_{j=1}^{m} \in \mathbb{R}^{m\times d}$. Here, n and m represent the number of unique codes in each dataset, respectively, and d denotes the dimension of the code embedding vector. We use these embedding tables to model the relationships between different codes using an attention mechanism. The attention score α_{ij} quantifies the importance of code j to code i. This score is calculated according to Equation 1:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[h_i \parallel h_j]))}{\sum_{k\in\mathcal{N}_i}\exp(\text{LeakyReLU}(a^T[h_i \parallel h_k]))} \quad (1)$$

where $h_i \in \mathbb{R}^d$ and $h_j \in \mathbb{R}^d$ represent the embedding vectors for codes i and j respectively, and $\|$ denotes concatenation operation. $a \in \mathbb{R}^{2d}$ is a learnable weight vector used to compute the attention coefficient between the two vectors. $\mathcal{N}_i$ denotes the set of codes in another system related to code i, including i itself. Finally, the code embedding vectors $u_i^{'}$ and $v_j^{'}$ used for training are calculated using Equation 2:

$$\begin{aligned} u_i^{'} &= \sigma(\alpha_{ii} u_i + \sum_{j \in \mathcal{N}_i \setminus \{i\}} \alpha_{ij} \omega_{ij} v_j) \\ v_j^{'} &= \sigma(\alpha_{jj} v_j + \sum_{i \in \mathcal{N}_j \setminus \{j\}} \alpha_{ji} \omega_{ji} u_i) \end{aligned} \quad (2)$$

where σ represents the ELU activation function and $\mathcal{N}_i \setminus \{i\}$ denotes the set of codes in another system related to code i, excluding i itself. In addition, weights are applied based on the approximate flag. If the codes i and j are completely equivalent, the weight ω_{ij} is set to 1. When they are approximately equivalent, the weight is set to 0.5. The calculated $U^{'} = \{u_i^{'}\}_{i=1}^{n} \in \mathbb{R}^{n \times d}$ and $V^{'} = \{v_j^{'}\}_{j=1}^{m} \in \mathbb{R}^{m \times d}$ are the final embedding tables for ICD-9 and ICD-10 codes, integrating information from both systems.

Building on this, the code connected through MAT can interact with code from another system during the update process, resulting in more refined and generalized code embeddings.

2.2. Training

JAMS retains the loss functions presented in each backbone model. However, unlike backbone models, which use a single dataset, JAMS uses datasets from both systems. Thus, if the datasets of the two systems are imbalanced within a batch, it leads to a bias for a particular system, which will negatively affect the model. To mitigate this problem, we introduce a normalized loss function in Equation 3:

$$\mathcal{L} = \frac{\mathcal{L}_{\text{ICD-9}} * \mu + \mathcal{L}_{\text{ICD-10}} * \eta}{\mu + \eta} \quad (3)$$

where μ and η denote the numbers of ICD-9 and ICD-10 data in a single batch. $\mathcal{L}_{\text{ICD-9}}$ represents the loss of data tagged as ICD-9 within a batch, while $\mathcal{L}_{\text{ICD-10}}$ is the loss of data tagged as ICD-10. $\mathcal{L}$ represents the overall batch loss. This normalized loss function helps the model learn from both coding systems equally without being affected by batch-level data imbalance.

3. Experiment

3.1. Experimental Settings

In this experiment, we used the MIMIC-IV ICD-9 and MIMIC-IV ICD-10 datasets. These datasets are based on MIMIC-IV (Johnson et al., 2023) and labeled according to their respective ICD versions. The statistics for each dataset are presented in Table 1. We evaluated the proposed method using the F1 score, AUC-ROC, Exact Match Rate (EMR), precision@k (k=8,15), R-precision, and Mean Average Precision (MAP) for comparison with the previous study (Edin et al., 2023).

	MIMIC-IV ICD-9	MIMIC-IV ICD-10
# of documents	209,326	122,279
# of patients	97,709	65,659
# of unique codes	6,150	7,942
Train / val / test [%]	73.8/10.5/15.7	72.9/10.9/16.2

Table 1: Statistics of MIMIC-IV ICD-9 and MIMIC-IV ICD-10 datasets

As our backbone models, we selected AMC models that adopt label-wise attention, specifically CAML (Mullenbach et al., 2018), MultiResCNN (Li and Yu, 2020), LAAT (Vu et al., 2020), and PLM-ICD (Huang et al., 2022). We then compared their performances after applying JAMS. For a fair comparison, we aligned the hyperparameters with the standards set by Edin et al. (2023) for each model. Furthermore, to demonstrate the stability and reliability of JAMS, we conducted experiments with three different seeds and reported the average results.

3.2. Main Results

Table 2 presents the experimental results for MIMIC-IV ICD-9 and MIMIC-IV ICD-10. JAMS improved the performance consistently regardless of the backbone models, demonstrating its model-agnostic characteristic, which is not constrained by specific model structures. In addition, the proposed approach enhanced the performance of both datasets. Particularly, as shown in Table 1, even though the MIMIC-IV ICD-10 has fewer documents and more unique codes than the MIMIC-IV ICD-9, our method showed a comparable increase in performance across both datasets. This result suggests effective knowledge transfer across the two coding systems, consistent with studies indicating that human coders familiar with ICD-9 perform better on ICD-10 tests (Sand and Elison-Bowers, 2013). Furthermore, our method showed more significant improvements in the macro-average scores, indicating its effectiveness in addressing the class imbalance issue.

	MIMIC-IV ICD-9										MIMIC-IV ICD-10									
	AUC-ROC		F1		EMR	Precision@k		R-precision	MAP	AUC-ROC		F1		EMR	Precision@k		R-precision	MAP
	Micro	Macro	Micro	Macro		8	15			Micro	Macro	Micro	Macro		8	15		
CAML	98.8	90.7	58.6	19.3	0.6	66.3	50.3	58.5	62.4	98.5	91.1	55.4	16.0	0.3	66.8	52.2	54.5	57.4
CAML†	**99.0**	**93.0**	**59.5**	**21.5**	**0.7**	**67.0**	**51.0**	**59.4**	**63.6**	**98.9**	**94.0**	**56.3**	**19.7**	0.3	**67.6**	**53.0**	**55.6**	**58.7**
MultiResCNN	99.2	95.1	60.4	27.7	0.8	67.6	51.8	60.4	64.7	99.0	94.5	56.9	21.1	0.4	67.8	53.5	56.1	59.3
MultiResCNN†	99.2	**96.0**	**60.9**	**29.2**	0.8	**68.0**	**52.2**	**61.0**	**65.4**	**99.1**	**96.2**	**57.9**	**23.5**	0.4	**68.9**	**54.4**	**57.3**	**60.9**
LAAT	99.3	96.0	61.7	26.4	0.9	68.9	52.7	61.7	66.3	99.0	95.4	57.9	20.3	0.4	68.9	54.3	57.2	60.6
LAAT†	99.2	95.7	**62.2**	**28.7**	**1.0**	**69.4**	**53.1**	**62.2**	**66.9**	**99.1**	**96.0**	**59.0**	**22.7**	**0.5**	**70.1**	**55.4**	**58.5**	**62.2**
PLM-ICD	99.4	97.2	62.6	29.8	1.0	70.0	53.5	62.7	68.0	99.2	96.6	58.5	21.1	0.4	69.9	55.0	57.9	61.9
PLM-ICD†	99.4	96.8	**62.8**	**31.8**	1.0	**70.2**	**53.8**	**63.1**	**68.3**	99.2	**96.7**	**59.8**	**25.6**	**0.5**	**71.0**	**56.1**	**59.3**	**63.6**

Table 2: Experimental results on the MIMIC-IV ICD-9 and MIMIC-IV ICD-10 test sets. Models marked with † indicate JAMS is applied.

3.3. Analysis

Performance Comparison by Code Frequency: We measured the performance based on the code frequency to assess the effectiveness of JAMS for low-frequency codes in an imbalanced label distribution setting. The codes were grouped into three categories (`1-100`, `101-1000`, `1001-`) based on their frequencies in the training data, and the macro F1 scores of LAAT and LAAT† were compared. The results are shown in Figure 2.

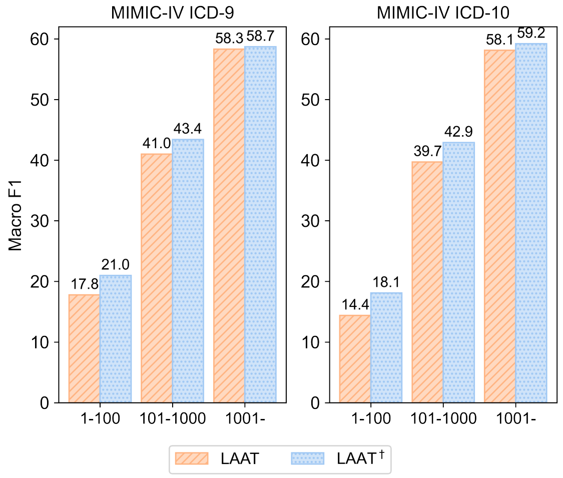

Figure 2: Performance comparison for groups based on code frequency.

As shown in Figure 2, LAAT† outperforms LAAT across all frequency groups in both datasets. Notably, the group with low-frequency codes (`1-100`) shows the most significant improvement, increasing by 3.2 in MIMIC-IV ICD-9 and 3.7 in MIMIC-IV ICD-10. These improvements are likely to be due to the effect of the generalization of low-frequency codes by leveraging related codes from another coding system. However, a group with high-frequency codes (`1001-`) could achieve sufficient generalization with their own system's data. Therefore, it can be speculated that the performance gains from incorporating data from another system are relatively limited.

Performance Comparison Based on Code Linkage: Building on the above results, we can confirm that JAMS effectively improves the performance of low-frequency codes. However, it remains unclear whether this is due to the linkages between different coding systems. To clarify this, we further divided the group with low-frequency codes (`1-100`) that showed the most significant performance improvement, based on whether they had corresponding mappings in the GEMs. Subsequently, we measured the performance using the macro F1 metric. Figure 3 shows the results of this comparison.

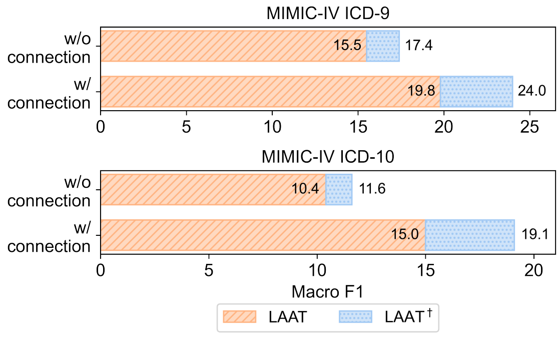

Figure 3: Performance comparison based on code linkage for low-frequency codes.

As shown in Figure 3, LAAT† achieves significant improvements in the group with connection information (`w/ connection`), with a gain of 4.2 in MIMIC-IV ICD-9 and 4.1 in MIMIC-IV ICD-10. These results indicate that our MAT-based code embedding method effectively generalizes low-frequency codes through connections across different coding systems. This finding provides significant evidence of the importance and effectiveness of leveraging relevance across these coding systems. Moreover, LAAT† shows improved performance even in the group without connection information (`w/o connection`), with an increase of 1.9 in MIMIC-IV ICD-9 and 1.2 in MIMIC-IV ICD-10. These performance gains are likely due to the shared encoder, which learns the medical documents of both datasets. This approach produces an effect similar to data augmentation and enhances the generalization ability of the encoder.

4. Conclusion

In this study, we introduce a novel approach to AMC called JAMS. Unlike most existing methods that consider only a single coding system, JAMS incorporates information from two different coding systems. It learns various representations and explicitly captures relationships across coding systems using MAT. Our experiments confirm that JAMS is model-agnostic and significantly improves the performance of low-frequency codes. In addition, our analysis shows that these performance gains are due to the connections between the codes of different systems. Based on our findings, we expect JAMS to be useful in various scenarios where different coding systems are used. These scenarios include migrations due to medical code version changes and transitions to different national coding systems. Therefore, our upcoming research will focus on expanding JAMS to a multilingual environment by incorporating the medical coding systems of two distinct countries.

5. Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2023-00216011, Development of artificial complex intelligence for conceptually understanding and inferring like human). Furthermore, this work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge)

6. Bibliographical References

Jinbo Bi, Tao Xiong, Shipeng Yu, Murat Dundar, and R. Bharat Rao. 2008. An improved multi-task learning approach with applications in medical diagnosis. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases(ECML/PKDD 2008)*, pages 117–132, Berlin, Heidelberg. Springer Berlin Heidelberg.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56, pages 301–318, Northeastern University, Boston, MA, USA. PMLR.

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(SIGIR '23)*, pages 2572–2582, New York, NY, United States. Association for Computing Machinery.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA. Association for Computational Linguistics.

Geunyeong Jeong, Juoh Sun, Seokwon Jeong, Hyunjin Shin, and Harksoo Kim. 2023. Improving automatic KCD coding: Introducing the KoDAK and an optimized tokenization method for Korean clinical documents. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 96–101, Toronto, Canada. Association for Computational Linguistics.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–9.

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.

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(5), pages 8180–8187, Palo Alto, California USA. AAAI Press.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Ramesh Kashyap, and Stefan Winkler. 2023. A two-stage decoder for efficient ICD coding. In *Findings of the Association for Computational Linguistics(ACL 2023)*, pages 4658–4665, Toronto, Canada. Association for Computational Linguistics.

Kimberly J O'Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: ICD code accuracy. *Health Services Research*, 40(5):1620–1639.

Jong Ku Park, Ki Soon Kim, Chun Bae Kim, Tae Yong Lee, Kang Sook Lee, Duk Hee Lee, Sunhee Lee, Sun Ha Jee, Il Suh, Kwang Wook Koh, et al. 2000. The accuracy of icd codes for cerebrovascular diseases in medical insurance claims. *Journal of Preventive Medicine and Public Health*, 33(1):76–82.

Jaime N Sand and Patt Elison-Bowers. 2013. ICD-10-CM/PCS: transferring knowledge from ICD-9-CM. *Perspectives in health information management*, 10(Summer):1g.

Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, and Alzheimer's Disease Neuroimaging Initiative. 2016. Deep sparse multi-task learning for feature selection in alzheimer's disease diagnosis. *Brain structure & function*, 221(5):2569–2587.

Wei Sun, Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2022. Multitask balanced and recalibrated network for medical code prediction. *ACM Trans. Intell. Syst. Technol.*, 14(1).

Lucia Otero Varela, Chelsea Doktorchik, Natalie Wiebe, Danielle A Southern, Søren Knudsen, Pallavi Mathur, Hude Quan, and Cathy A Eastwood. 2022. International classification of diseases clinical coding training: An international survey. *Health Information Management Journal*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. *International Conference on Learning Representations*.

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, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. A survey of automated International Classification of Diseases coding: development, challenges, and applications. *Intelligent Medicine*, 2(3):161–173.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. 2023. Multi-label few-shot ICD coding as autoregressive generation with prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(4), pages 5366–5374, Washington, DC, USA. AAAI Press.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland. Association for Computational Linguistics.

Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. 2022. Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2883–2891, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.