Bridging the Version Gap: Multi-version Training Improves ICD Code Prediction, Especially for Rare Codes

Jinghui Liu **Anthony Nguyen**
Australian e-Health Research Centre, CSIRO
{jinghui.liu,anthony.nguyen}@csiro.au

Abstract

Clinical coding maps clinical documentation to standardized medical codes, an essential yet time-consuming administrative task that could benefit from automation. Current models on ICD coding are typically optimized for codes from a specific ICD version. However, in reality, ICD systems evolve continuously, and different versions are adopted across time periods and regions. Moreover, ICD coding suffers from the long-tail problem, and rare code performance can be a bottleneck for developing implementable models. We examine whether it is viable to train version-independent models by combining data annotated in different ICD versions, which may help address these challenges. We add ICD-9 data to the training of a modified label-wise attention model for ICD-10 prediction, and find that despite the version mismatch, adding ICD-9 yields a 27% increase in micro F1 for 18K rare ICD codes compared to training on ICD-10 alone. On 8K frequent ICD-10 codes, the multi-version training also substantially improves macro metrics, with far fewer model parameters.

1 Introduction

The International Classification of Diseases (ICD) is the global standard for reporting health conditions and diseases. It plays a critical role in healthcare billing, epidemiological research, and health policymaking (Dong et al., 2022; Gan et al., 2025). Assigning ICD codes to clinical documents requires a high level of expertise given that the candidate code set is often large (e.g., ICD-10 includes over 70K diagnosis codes) and the documentation can be lengthy and complex (Motzfeldt et al., 2025; Liu et al., 2023). Consequently, the coding process is time-consuming even for experienced coders, which has motivated extensive research into NLP models that automate or assist the coding process (Stanfill et al., 2010; Ji et al., 2024).

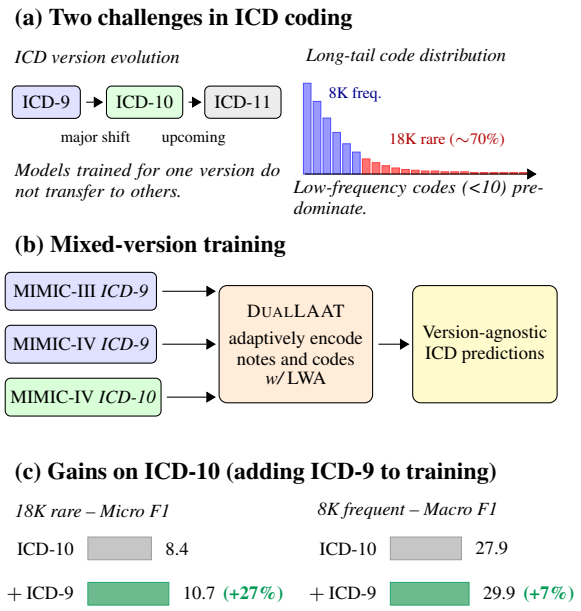

Figure 1: **(a)** ICD coding faces two intertwined challenges: the ICD system evolves continuously and the code distribution is heavily long-tailed. **(b)** We mix three MIMIC-derived datasets spanning ICD-9 and ICD-10 to train a single version-agnostic model, DUALLAAT. **(c)** Adding ICD-9 to ICD-10 training yields a 27% relative gain in micro F1 on rare ICD-10 codes than training on ICD-10 alone.

Current best-performing ICD coding models typically frame the task as a multi-label, multi-class classification problem (Douglas et al., 2025; Edin et al., 2023; Yuan et al., 2022), which requires defining a fixed label space and a substantial amount of data for supervised training. This setup faces two critical challenges in real-world ICD coding. First, **the ICD coding system undergoes continuous updates**, with new codes added and old codes retired. For example, CMS actively updates ICD-10 every year to keep codes aligned with healthcare needs.[1] At certain points, the ICD system undergoes major

[1] https://www.cms.gov/medicare/coding-billing/icd-10-codes

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 372–381

transitions, such as the shift from ICD-9 to ICD-10, and the anticipated move to ICD-11 (Harrison et al., 2021). Models with a pre-defined label space have trouble handling these updates, which introduce substantial data shifts (Finlayson et al., 2021).

Second, **ICD code distributions are long-tailed** (Li et al., 2025a; He et al., 2025), which is a critical modeling challenge as the ICD system grows larger and more fine-grained. For example, in MIMIC-IV (Johnson et al., 2023), infrequent ICD-10 codes that appear fewer than 10 times (18K) account for 69.6% of all codes in the database (26K). Collecting sufficient samples for these rare and underrepresented diseases is essential for training. However, achieving this requires consolidating patient records across time periods and regions, which is likely accompanied by mismatched ICD versions. It remains understudied whether such a combination leads to benefits or harm in modeling. For example, given that only 24.3% of ICD-9-CM codes have exact matches in ICD-10-CM (Fung et al., 2021), mixing multiple datasets across versions may plausibly introduce noise that harms rather than helps modeling.

This study explores this question by examining if ICD-9 data is valuable for training an ICD-10 prediction model (Figure 1). Our experiments used two ICD-9 datasets and one ICD-10 dataset based on MIMIC (Edin et al., 2023), and mixed their training sets to train a version-independent model based on label-wise attention (LWA) (Wu et al., 2024; Mullenbach et al., 2018). Adding ICD-9 data to the training set led to **substantially improved performance on rare codes** (micro F1: $8.4 \rightarrow 10.7$) in ICD-10 using **a much smaller model**, as well as **improved per-label performance on frequent codes** (macro F1 of 29.9). Our findings demonstrate the synergistic effect of combining multi-version datasets for ICD coding, showcasing the potential value of leveraging legacy patient records to train more robust ICD prediction systems.

2 Related Work

Research on automated and assisted ICD coding dates back to the 1990s (Larkey and Croft, 1996). Recent advances have largely focused on configuring neural architectures to better model the large code set (Li et al., 2025b; Shi et al., 2017). Ji et al. (2024) surveyed modeling approaches and highlighted that label-wise attention (LWA) (Mullenbach et al., 2018; Wu et al., 2024; Vu et al., 2020) remains a key module to train high-performing models. Edin et al. (2023) conducted a comprehensive benchmark of various LWA-based models across three datasets spanning MIMIC-III and IV and ICD-9 and ICD-10, finding that PLM-ICD with in-domain pretrained checkpoints performed best. Other approaches exploit the relations between codes in the ICD ontology to improve modeling (Luo et al., 2024; Yuan et al., 2022). To the best of our knowledge, all prior models were trained for a specific set of ICD codes as label space, and none were evaluated across different ICD versions.

A growing number of studies have explored prompting LLMs to predict ICD codes (Simmons et al., 2025; Yang et al., 2023). While simple prompting achieves low performance (Soroush Ali et al., 2024), agent-based systems for ICD prediction show promise (Motzfeldt et al., 2025; Zheng et al., 2025). The generative approach is not constrained by a fixed label space and may offer enhanced interpretability. However, most of these studies were conducted on small-scale datasets with a limited label space of about 1K (Cheng et al., 2023), which contrasts with existing benchmarks (8K in Edin et al. (2023)) and public databases (26K in Johnson et al. (2023)). In addition, agent-based approaches often require large token budgets, which can be costly. For this study, we focus on comparing with supervised approaches on datasets with large code sets.

3 Methods and Experiments

3.1 Benchmark Datasets

We adopted the benchmark from Edin et al. (2023), which includes three MIMIC-derived datasets, namely MIMIC-III (ICD-9), MIMIC-IV (ICD-9), and MIMIC-IV (ICD-10). The original benchmark focuses on frequent codes, defined as appearing $\geq$ 10 times in the corresponding database. To also evaluate rare codes, we re-extracted all codes from MIMIC, resulting in substantial expansions of the code space ($6K \rightarrow 11K$ for ICD-9 and $8K \rightarrow 26K$ for ICD-10). Since ICD-9 is no longer in active use, we focused on ICD-10 as the primary evaluation benchmark. Dataset statistics are presented in Appendix Table 5.

3.2 Classification Model

LWA trains code-specific embeddings to attend to relevant text tokens in clinical notes and uses them for classification (Edin et al., 2024). To enable

adaptive modeling across ICD versions, we trained an encoder to generate code embeddings from code descriptions, which are unique textual strings describing each ICD code. We jointly trained two text encoders – one for clinical notes and one for ICD codes – using LWA. Given the dual-encoder feature and the use of LWA from LAAT (Vu et al., 2020), we denote the model as DUALLAAT. The model takes N notes and C codes (as text strings) as input and outputs a probability matrix $\hat{\mathbf{y}} \in \mathbb{R}^{N \times C}$ as prediction. The codes in C are version-agnostic and can be ICD-9 or ICD-10. Detailed notations and training procedures are provided in Appendix B.

3.3 Experiments

To investigate the impact of mixing multiple ICD data sources, we aggregated the three train sets from MIMIC-III (ICD-9), MIMIC-IV (ICD-9), and MIMIC-IV (ICD-10) for training and evaluated on the corresponding test sets. We included all ICD codes (i.e., from full label space) in the aggregated data to train the model. To enable fast experimentation and to highlight the impact of data mixing, we used basic CNN (Mullenbach et al., 2018) and RNN (Vu et al., 2020) as text encoders, denoted as $\text{DUALLAAT}_{\text{cnn}}$ and DUALLAAT, respectively.

We adopted the benchmark metrics (Edin et al., 2023) for evaluation on frequent codes, including both classification and ranking metrics. In particular, macro F1 was calculated as the arithmetic mean of F1 scores per code. Higher macro F1 indicates stronger performance on less frequent codes, reflecting tail-end robustness. We also evaluated on full ICD-10 set and explicitly on rare codes, which in practice can be equally important as frequent codes for reporting disease prevalence and enabling appropriate reimbursement.

3.4 Baselines

We consider two strong supervised models as baselines for frequent and full code sets, respectively:

PLM-ICD. This model adopts in-domain BERT with sliding-window as the text encoder and trains with LWA (Huang et al., 2022). Prior work found replacing BERT with Llama on a dataset with 50 codes did not improve performance (Motzfeldt et al., 2025), highlighting the strong baseline performance of PLM-ICD. We report its results from the previous benchmark study on the frequent 8K codes (Edin et al., 2023).

CoRelation. This model improves code representation learning by modeling the relations between codes through a bipartite graph for enhanced accuracy and efficiency (Luo et al., 2024). It was evaluated on top-50 and full codes in MIMIC. We use this model for reference comparison on full ICD-10 set with 26K codes.

4 Results

4.1 Impact of Data Mixing on Frequent Code Prediction and Benchmark Comparison

Table 1 presents the results on frequent codes. On ICD-10 prediction, mixing additional ICD-9 datasets from MIMIC improves all metrics for both $\text{DUALLAAT}_{\text{cnn}}$ and $\text{DUALLAAT}_{\text{rnn}}$, with about 2 points of improvement on macro F1. This shows that adding ICD-9 to training is beneficial for ICD-10 prediction despite terminological differences. Using a stronger encoder (RNN over CNN) also brings steady improvements, consistent with prior findings (Edin et al., 2023) and suggesting potential gains with stronger encoder models.

Compared with the PLM-ICD baseline, RNN-based DUALLAAT trained on mixed MIMIC datasets achieved similar precision but much higher macro AUC-ROC and macro F1. We also report results for the weaker CAML (Mullenbach et al., 2018) and LAAT (Vu et al., 2020) in Appendix Table 6. These baseline methods with CNN and RNN encoders form a more direct comparison with our model. The additional results also show that DUALLAAT trained using mixed datasets offers lower variance across the metrics.

4.2 Impact on Rare Code Prediction

Mixing additional data sources shows substantial benefits in rare code prediction, as presented in Table 2. Both ICD-10 and 9 benefited from alternative ICD data. For ICD-10, this yielded a 27.4% increase in micro F1, which is a substantial improvement given the number of codes (18K).

Table 4 compares the model performance on the full code set with 26K labels from MIMIC-IV ICD-10. DUALLAAT achieves results comparable to the reported CoRelation numbers on shared metrics, with a notably higher macro F1 (6.3→11.2), suggesting stronger performance across the long tail of rare codes. This shows that enhancing training data can be as important in ICD coding as architectural improvements, which have been the predominant focus in ICD prediction literature (Ji et al., 2024; Dong et al., 2022).

	Classification				Ranking			
	AUC-ROC		F1		Precision@k		R-precision	MAP
	Micro	Macro	Micro	Macro	8	15		
MIMIC-III ICD-9								
PLM-ICD (Edin et al., 2023)	98.9	95.9	59.6	26.6	72.1	56.5	60.1	64.6
DUALLAAT	**99.2**	**96.9**	**61.8**	**33.3**	**74.9**	**58.8**	**62.8**	**68.1**
MIMIC-IV ICD-9								
PLM-ICD (Edin et al., 2023)	99.4	97.2	62.6	29.8	70.0	53.5	62.7	68.0
DUALLAAT	**99.5**	**97.6**	**63.4**	**34.3**	**71.0**	**54.3**	**63.8**	**69.3**
MIMIC-IV ICD-10								
PLM-ICD (Edin et al., 2023)	99.2	96.6	**58.5**	21.1	**69.9**	**55.0**	57.9	61.9
DUALLAAT$_{cnn}$ (ICD-10 only)	99.2	96.8	55.6	24.3	67.5	52.7	55.3	59.0
DUALLAAT$_{cnn}$	99.3 (↑)	97.2 (↑)	56.6 (↑)	26.0 (↑)	68.4 (↑)	53.6 (↑)	56.4 (↑)	60.4 (↑)
DUALLAAT (ICD-10 only)	99.3	97.1	57.5	27.9	69.3	54.4	57.2	61.5
DUALLAAT	**99.3** (–)	**97.4** (↑)	58.0 (↑)	**29.9** (↑)	**69.9** (↑)	54.9 (↑)	**58.0** (↑)	**62.3** (↑)

Table 1: Results on frequent ICD codes for MIMIC-III *ICD-9* (3.7K), MIMIC-IV *ICD-9* (6K) and MIMIC-IV *ICD-10* (8K). The means of three random seed runs were reported. For ICD-10 prediction, ↑ indicates improvement of multi-source training (ICD-10 + ICD-9) compared to training on ICD-10 alone.

4.3 Training and Parameter Efficiency

Mixing multiple ICD sources for training also improves efficiency as the model can now be applied to datasets with different ICD versions, in effect reducing the number of model parameters required. In the context of this study, three PLM-ICD models are needed for the three benchmarks on frequent code prediction, whereas one DUALLAAT can handle them all. Training with mixed data also shortens convergence time: our training plateaued at around 10 epochs compared to 20 epochs for PLM-ICD.

Table 3 reports the comparison between PLM-ICD and DUALLAAT. While DUALLAAT with mixed datasets achieves comparable performance, it uses only a fraction of the training time and parameters. This can be an important factor to consider when deploying a model in a low-resource hospital or outpatient setting (Wu et al., 2022).

	F1 Micro	Precision@8	MAP
MIMIC-IV-ICD10			
Single Dataset	8.4	5.8	19.5
+ Mixing Alt ICD	10.7 (27.4% ↑)	6.1 (5.2% ↑)	21.4 (9.7% ↑)
MIMIC-IV-ICD9			
Single Dataset	7.2	5.6	25.4
+ Mixing Alt ICD	10.9 (51.4% ↑)	6.6 (17.9% ↑)	31.2 (22.8% ↑)

Table 2: Results on the rare codes (frequency < 10 in the cohort). DUALLAAT$_{cnn}$ trained on the corresponding train set was used as baseline.

	Training time	# Param
ICD-10 only		
PLM-ICD	34 hrs	137M
DUALLAAT$_{cnn}$	5 hrs	15M
DUALLAAT	18 hrs	37M
ICD-10 + ICD-9 (×3 datasets)		
PLM-ICD (×3 models)	95 hrs	402M
DUALLAAT$_{cnn}$	17 hrs	15M
DUALLAAT	57 hrs	37M

Table 3: Comparison of training time and model size. Mixed training data includes one dataset for ICD-10 and two datasets for ICD-9.

5 Discussion & Conclusion

Training effective ICD coding models is challenging due to the long-tail distribution of codes and the continuous updates to the ICD system. Most existing work focuses on improving modeling methods, yet little attention has been paid to the impact of the training data itself. We show that mixing multiple data sources – even with varied ICD versions – benefits both frequent and rare code predictions. The gains are notable given that less than 25% of ICD-9 have equivalent mappings to ICD-10 (Fung et al., 2021), suggesting the benefits arise from shared clinical semantics rather than direct label overlap, which warrants future investigation.

The viability of mixing ICD versions means **legacy patient records can be utilized in training new models**, an important implication for developing coding systems to handle ongoing modifica-

	AUC-ROC Micro	AUC-ROC Macro	F1 Micro	F1 Macro	Precision@8
PLM-ICD (Luo et al., 2024)	99.0	91.9	57.0	4.9	69.5
CoRelation (Luo et al., 2024)	99.6	97.2	57.8	6.3	70.0
DUALLAAT (ICD-10 only)	99.6	97.0	56.6	10.2	69.3
DUALLAAT	99.7	97.4	57.1	11.2	69.8

Table 4: Results on 26K ICD-10 codes in MIMIC-IV. PLM-ICD and CoRelation results are from the original paper, which uses a different preprocessing pipeline; we include this approximate comparison for reference.

tions of ICD rules. For major rule updates, a model trained on legacy records can avoid cold-start issues and serve as a foundation for fine-tuning.

In conclusion, we demonstrate the advantages of mixing multi-source datasets to train ICD prediction models, resulting in more accurate, robust, and efficient performance. The synergistic effect of combining diverse sources demonstrates the potential for further scaling both training data and model capacity to advance the progress of ICD coding research. Finally, our code and model checkpoints are released to support reproducibility.[2]

Limitations

ICD and Data Coverage. This study only considered the MIMIC databases as the data sources, which originate from a single institution in the United States and use ICD Clinical Modification (CM). Other regional implementations of ICD exist, such as ICD-10-CA for Canada and ICD-10-AM for Australia. Furthermore, many countries adopt ICD systems in non-English languages, including China (CCD/ICD-10-CN) and Germany (ICD-10-GM). We were limited by the scope of data in this initial study and focused on openly available datasets. Given insights from multilingual language modeling, cross-lingual knowledge transfer (Artetxe et al., 2020) may also be feasible for ICD modeling. However, this warrants future research, particularly to address nuanced differences between different ICD versions beyond mere translation and semantic mappings.

Classification Model. For this study, we employed simple classification models in the experiments, but we believe the patterns of data mixing can be extrapolated to other models given the existing findings on modeling architectures (Ji et al., 2024; Edin et al., 2023). Training PLM-based models with multi-source datasets represents a natural follow-up, though we leave this open as newer and more effective modeling methods may emerge in the near future.

Record Overlap. An inherent challenge in mixing data sources is potential overlapping records. MIMIC-III (2001-2012) and MIMIC-IV (2008-2018) overlap temporally, but we could not explicitly identify shared patients because they use different ID conventions and are not linkable. This may have influenced results for ICD-9, although these were not the primary focus of our experiments. Meanwhile, this is less of an issue for ICD-10 since it is distinctly different from ICD-9 and only exists in MIMIC-IV. We found seven patients appearing in both ICD-9 and ICD-10 datasets from Edin et al. (2023) in MIMIC-IV, with only one patient in the ICD-10 test set (19,802 patients). Given the large test size and different data distributions, we deemed it acceptable to retain this patient to enable direct comparison with the baselines.

[2]The training code and model checkpoints can be found in https://github.com/JHLiu7/Dual-LAAT.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159.

James C Douglas, Yidong Gan, Ben Hachey, and Jonathan K Kummerfeld. 2025. Less is more: Explainable and efficient ICD code prediction with clin-

ical entities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30835–30847, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, USA. Association for Computing Machinery.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Stroudsburg, PA, USA. Association for Computational Linguistics.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. 2021. The clinician and dataset shift in artificial intelligence. *The New England journal of medicine*, 385(3):283–286.

Kin Wah Fung, Julia Xu, Shannon McConnell-Lamptey, Donna Pickett, and Olivier Bodenreider. 2021. Feasibility of replacing the ICD-10-CM with the ICD-11 for morbidity coding: A content analysis. *Journal of the American Medical Informatics Association*, 28(11):2404–2411.

Yidong Gan, Maciej Rybinski, Ben Hachey, and Jonathan K Kummerfeld. 2025. Aligning AI research with the needs of clinical coding workflows: Eight recommendations based on US data analysis and critical review. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 909–922, Stroudsburg, PA, USA. Association for Computational Linguistics.

James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. 2021. ICD-11: an international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making*, 21(Suppl 6):206.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *An International Journal on Information Fusion*, 118(102963):102963.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shaoxiong Ji, Xiaobo Li, Wei Sun, Hang Dong, Ara Taalas, Yijia Zhang, Honghan Wu, Esa Pitkänen, and Pekka Marttinen. 2024. A unified review of deep learning for automated medical coding. *ACM computing surveys*.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, 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.

Leah S Larkey and W Bruce Croft. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 289–297, New York, NY, USA. Association for Computing Machinery.

Rumeng Li, Xun Wang, and Hong Yu. 2025a. Improving rare and common ICD coding via a multi-agent LLM-based approach. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4945–4949, New York, NY, USA. ACM.

Xiaobo Li, Yijia Zhang, Xiaodi Hou, Shilong Wang, and Hongfei Lin. 2025b. Deep learning for automatic ICD coding: Review, opportunities and challenges. *Artificial intelligence in medicine*, 168(103187):103187.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2023. Automated ICD coding using extreme multi-label long text transformer-based models. *Artificial intelligence in medicine*, 144(102662):102662.

Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. 2024. CoRelation: Boosting automatic ICD coding through contextualized code relation learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3997–4007.

Andreas Geert Motzfeldt, Joakim Edin, Casper L Christensen, Christian Hardmeier, Lars Maaløe, and Anna Rogers. 2025. Code like humans: A multi-agent solution for medical coding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22612–22627, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated ICD coding using deep learning. *arXiv [cs.CL]*.

Ashley Simmons, Kullaya Takkavatakarn, Megan McDougal, Brian Dilcher, Jami Pincavitch, Lukas Meadows, Justin Kauffman, Eyal Klang, Rebecca Wig, Gordon Smith, Ali Soroush, Robert Freeman, Donald J Apakama, Alexander W Charney, Roopa Kohli-Seth, Girish N Nadkarni, and Ankit Sakhuja. 2025. Extracting international classification of diseases codes from clinical documentation using large language models. *Applied clinical informatics*, 16(2):337–344.

Soroush Ali, Glicksberg Benjamin S., Zimlichman Eyal, Barash Yiftach, Freeman Robert, Charney Alexander W., Nadkarni Girish N, and Klang Eyal. 2024. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040.

Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association: JAMIA*, 17(6):646–651.

Ashish Vaswani, Noam M Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Information Processing Systems*, 30:5998–6008.

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. International Joint Conferences on Artificial Intelligence Organization.

Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael T C Poon, Natalie Fitzpatrick, Adam P Levine, Luke T Slater, Alex Handy, Andreas Karwath, Georgios V Gkoutos, Claude Chelala, Anoop Dinesh Shah, Robert Stewart, Nigel Collier, Beatrice Alex, and 4 others. 2022. A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *NPJ digital medicine*, 5(1):186.

John Wu, David Wu, and Jimeng Sun. 2024. Beyond label attention: Transparency in language models for automated medical coding via dictionary learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8848–8871, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhichao Yang, Sanjit Singh Batra, Joel Stremmel, and Eran Halperin. 2023. Surpassing GPT-4 medical coding with a two-stage approach. *arXiv [cs.CL]*.

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.

Jiyang Zheng, Islam Nassar, Thanh Vu, Xu Zhong, Yang Lin, Tongliang Liu, Long Duong, and Yuan-Fang Li. 2025. MedDCR: Learning to design agentic workflows for medical coding. *arXiv [cs.AI]*.

A Datasets for Evaluation

Table 5 presents the statistics of the three benchmark datasets from Edin et al. (2023). We additionally extracted the infrequent codes that were not included in the prior study to evaluate the models' robustness on tail-end ICD codes.

B Classification Model in Detail

B.1 DUALLAAT with dual encoders

Given a clinical note N with t_{note} word tokens $w_1^N, w_2^N, \ldots, w_{t_{\text{note}}}^N$, we have a clinical code C with t_{code} word tokens $w_1^C, w_2^C, \ldots, w_{t_{\text{code}}}^C$ from the code description. They are transformed into word embeddings by a shared embedding layer into $\mathbf{e}_1^N, \mathbf{e}_2^N, \ldots, \mathbf{e}_{t_{\text{note}}}^N$ and $\mathbf{e}_1^C, \mathbf{e}_2^C, \ldots, \mathbf{e}_{t_{\text{code}}}^C$, respectively.

Clinical note representations at token level $\mathbf{H}_{\text{note}} \in \mathbb{R}^{d_{\text{note}} \times t_{\text{note}}}$ are created by a note encoder as $encoder_{note}$:

$$\mathbf{H}_{\text{note}} = \text{encoder}_{\text{note}}(\mathbf{e}_1^N, \ldots, \mathbf{e}_{t_{\text{note}}}^N) \quad (1)$$

The clinical code representation is encoded by a separate $encoder_{code}$ and pooled into a code-level representation. The representation for one code $\mathbf{h}_{\text{code}} \in \mathbb{R}^{d_{\text{code}}}$ is created as:

$$\mathbf{h}_{\text{code}} = \text{Pooling}(\text{encoder}_{\text{code}}(\mathbf{e}_1^C, \ldots, \mathbf{e}_{t_{\text{code}}}^C)) \quad (2)$$

In practice, we consider L codes by encoding them in a batch to create $\mathbf{H}_{\text{code}} \in \mathbb{R}^{|L| \times d_{\text{code}}}$. Then we have two projection matrices $\mathbf{W}_{\text{note}} \in \mathbb{R}^{d_{\text{shared}} \times d_{\text{note}}}$ and $\mathbf{W}_{\text{code}} \in \mathbb{R}^{d_{\text{code}} \times d_{\text{shared}}}$ for note and code, respectively. The projected code representations $\mathbf{H}_{\text{code}} \cdot \mathbf{W}_{\text{code}} \in \mathbb{R}^{|L| \times d_{\text{shared}}}$ replaces the

	MIMIC-III *ICD-9*	MIMIC-IV *ICD-9*	MIMIC-IV *ICD-10*
Number of notes	52,712	209,326	122,278
Notes: Train/val/test [%]	72.9/10.6/16.6	73.8/10.5/15.7	72.9/10.9/16.2
	Frequent Code Set		
Number of unique codes	3,681	6,150	7,942
Codes per note: Median (IQR)	14 (10-20)	12 (8-17)	14 (9-20)
Codes per note: Mean (Std)	15.6±8.0	13.3±7.6	15.7±8.7
	Full Code Set		
Number of unique codes	8,925	11,324	26,085
Codes per note: Median (IQR)	14 (10-20)	12 (8-17)	15 (10-21)
Codes per note: Mean (Std)	15.9±8.1	13.4±7.6	16.0±8.9
	Rare Code Set		
Number of unique codes	5,244	5,174	18,143
Codes per note: Median (IQR)	1 (1-2)	1 (1-1)	1 (1-2)
Codes per note: Mean (Std)	1.5±1.1	1.2±0.6	1.6±1.2
Number of notes (at least 1 rare code)	10,653	14,185	30,332

Table 5: Benchmark datasets on MIMIC-III (ICD-9), MIMIC-IV (ICD-9), and MIMIC-IV (ICD-10) from Edin et al. (2023), which are based on the *frequent code set*. Additional *full code set* and *rare code set* are also examined. Rare codes refer to codes occurring less than 10 times in the dataset, and were not included in Edin et al. (2023). *Full code set* share the same clinical notes with the *frequent code set*. IQR: interquartile range. Std: standard deviation.

code-specific embeddings in typical LWA implementation. With an additional dimension d_{shared} specified, the attention scores and representations based on dual encoders are computed:

$$\mathbf{A}_{\text{dual}} = \text{softmax}\Big(\tanh(\mathbf{H}_{\text{code}} \cdot \mathbf{W}_{\text{code}}) \cdot \tanh(\mathbf{W}_{\text{note}} \cdot \mathbf{H}_{\text{note}}) \Big) \quad (3)$$

$$\mathbf{J}_{\text{dual}} = \mathbf{H}_{\text{note}} \cdot \mathbf{A}_{\text{dual}}^{\top} \quad (4)$$

This formulates the backbone of DUALLAAT, which use separate note encoder and code encoder to create their corresponding representations and compute label-wise attention.

To further enhance the dual label-wise attention, we also consider M attention heads with $\mathbf{W}^1_{\text{note}}, \ldots, \mathbf{W}^M_{\text{note}}$ and $\mathbf{W}^1_{\text{code}}, \ldots, \mathbf{W}^M_{\text{code}}$ as multi-head attention (MHA) (Vaswani et al., 2017). Then we compute $\mathbf{J}^1_{\text{dual}}, \ldots, \mathbf{J}^M_{\text{dual}}$ and obtain the multi-head note representation per code by concatenation:

$$\mathbf{J}^{\text{mha}}_{\text{dual}} = \text{concat}(\mathbf{J}^1_{\text{dual}}, \ldots, \mathbf{J}^M_{\text{dual}}) \quad (5)$$

The final output is computed as:

$$\hat{\mathbf{y}} = \text{sigmoid}(\text{classifier}(\mathbf{J}^{\text{mha}}_{\text{dual}})) \quad (6)$$

The model is trained to minimize binary cross entropy loss. Different from the previous models that consider only note N as input, DUALLAAT now takes both N (notes) and C (codes) as inputs. This enables a flexible choice of L by changing C either during training or at inference. Since the clinical codes C are provided as code descriptions in text, it allows flexibility to feed different versions of ICD codes as long as they are provided in the textual form. During inference, these descriptions can be directly mapped back to the actual codes.

B.2 Training DUALLAAT

To train DUALLAAT in batches, we need to dynamically set the label space L for different codes. For example, given a batch of 32 patients with each patient having 20 codes on average, L would be set dynamically according to appearing clinical codes with $|L| = 640$ (20×32). Meanwhile, in reality the number of clinical codes can vary dramatically from patient to patient, resulting in large variations in code numbers that would destabilize the label space and thus training. In addition, patients in a batch sometimes may contain codes in overlap, in effective reducing the scope of L.

To address this, we specify the label space L to be a fixed number that is larger than the possible codes to appear in a batch. Given the fixed L, we collect the unique codes in the batch as positives L_{pos}, and then randomly sample negative codes L_{neg} from a code pool to form L so that $|L| = |L_{pos}| + |L_{neg}|$. This setup has three advantages. First, it fixes the number of L to stabilize training. Second, it removes duplicated positive codes and allows negative sampling to expand the scope of L, enabling higher sample efficiency. Finally, $|L|$ becomes a hyperparameter to be controlled.

During training, we use a sampler to keep only

one ICD version in each batch, i.e., L is either L^{icd9} or L^{icd10}. These batches of different ICD versions are mixed and shuffled to train the model with standard mini-batch gradient descent.

B.3 Hyperparameter setting

We set the dimensions of the encoders and the projection layers the same $d_{note} = d_{code} = d_{shared}$ without further tuning. We follow the implementation from (Edin et al., 2023) as best practice in setting the hyperparameters, where DUALLAAT has a dimension of 512, one bidirectional layer, and dropout rate of 0.3; and DUALLAAT$_{\text{cnn}}$ has a 256 filters with filter width of 10, and dropout rate of 0.2. All models are trained using a linear scheduler with learning rate of 0.001 and 2000 warmup steps on batches of 32 samples. DUALLAAT is additionally imposed with weight decay of 0.001 to regularize training. The maximum number of tokens in clinical notes is set to 4000, and that for the clinical code descriptions is set to 48.

We pretrain the word embeddings (100 dimensions) based on the three train sets, again following the setup in (Edin et al., 2023). By default, we train the models on full ICD codes, and set the label space size $|L|$ to 8192. We use GRU as the default RNN choice for DUALLAAT$_{\text{rnn}}$, with an alternative DUALLAAT$_{\text{cnn}}$ using CNN.

	CAML	LAAT	PLM-ICD	DUALLAAT$_{cnn}$ (+ ICD-9)	DUALLAAT (+ ICD-9)
AUC-ROC Micro	98.5 ± 0.0	99.0 ± 0.1	99.2 ± 0.0	99.3 ± 0.0	$\mathbf{99.3 \pm 0.0}$
AUC-ROC Macro	91.1 ± 0.1	95.4 ± 0.3	96.6 ± 0.2	97.2 ± 0.0	$\mathbf{97.4 \pm 0.0}$
F1 Micro	55.4 ± 0.2	57.9 ± 0.1	$\mathbf{58.5 \pm 0.7}$	56.6 ± 0.2	58.0 ± 0.0
F1 Macro	16.0 ± 0.3	20.3 ± 0.4	21.1 ± 2.3	26.0 ± 0.2	$\mathbf{29.9 \pm 0.3}$
P@8	66.8 ± 0.2	68.9 ± 0.1	$\mathbf{69.9 \pm 0.6}$	68.4 ± 0.2	$\mathbf{69.9 \pm 0.1}$
P@15	52.2 ± 0.1	54.3 ± 0.1	$\mathbf{55.0 \pm 0.6}$	53.6 ± 0.2	54.9 ± 0.1
R-precision	54.5 ± 0.2	57.2 ± 0.1	57.9 ± 0.8	56.4 ± 0.2	$\mathbf{58.0 \pm 0.1}$
MAP	57.4 ± 0.2	60.6 ± 0.2	61.9 ± 0.9	60.4 ± 0.2	$\mathbf{62.3 \pm 0.1}$

Table 6: Results on frequent ICD-10 codes (main benchmark) with additional baseline models. Scores are reported with standard deviation.