MIMIC-IV-ICD:
A New Benchmark for Multi-Label Medical Code Classification

Anonymous ACL submission

Abstract

Clinical notes are assigned ICD codes – sets of codes for diagnoses and procedures. In recent years, predictive machine learning models have been built for automatic ICD coding. However, there is a lack of widely accepted benchmarks for automated ICD coding models based on large-scale public EHR data. This paper proposes a public benchmark suite for ICD-10 coding using a large EHR dataset derived from MIMIC-IV, the most recent public EHR dataset. We standardize data preprocessing and establish a comprehensive ICD coding benchmark dataset. Some state-of-the-art models for ICD prediction are thoroughly investigated, and we provide benchmark results as useful references for future studies. Our open-source code offers easy access to data processing steps, benchmark creation, and experiment replication for those with MIMIC-IV access, providing insights, guidance, and protocols to efficiently develop ICD coding models.

1 Introduction

Medical records and clinical documentation serve as essential sources of information on patient care, disease progression, and healthcare operations. Following a patient's visit, medical coders analyze these records and identify diagnoses and procedures according to the International Classification of Diseases (ICD) system (WHO, 1948). These codes are used in predictive modeling for patient care and health status, as well as for insurance claims, billing processes, and other hospital functions (Tsui et al., 2002).

Despite healthcare advancements, manual ICD coding remains problematic. This task, involving the interpretation of detailed medical records and the selection of appropriate codes from an extensive list (18,000 in ICD-9 and 155,000 in ICD-10), is daunting. Coders scrutinize medical notes to identify phrases aligning with code descriptions, such as associating "viral gastroenteritis" with the code A00.84 for "viral intestinal infection". This tedious process risks coding errors, potentially causing financial loss or resource misallocation in patient care. Hence, automated ICD coding is gaining attention from both industry and academia.

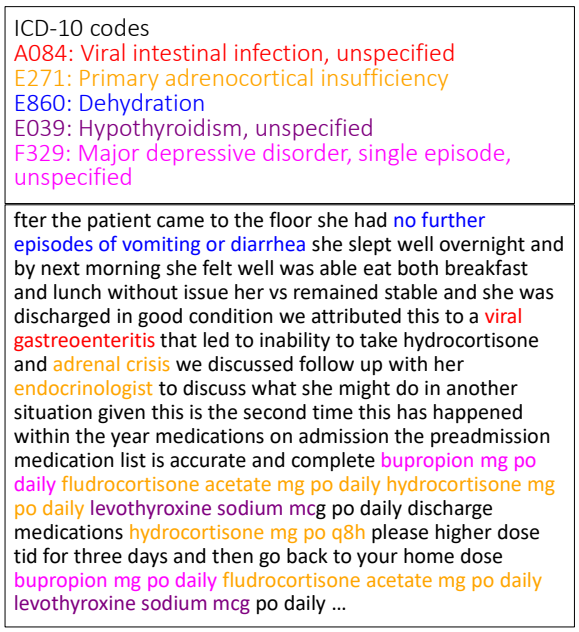

Figure 1: An example of an EHR note with ICD-10 codes (top) matched with its related discharge note text (bottom) from the MIMIC-IV dataset. For clarity, each code and its related mentions or evidence within the note text are color-coded.

While a variety of machine learning approaches have been proposed for ICD prediction (see a survey in Section 3), a significant limitation is their singular reliance on the MIMIC-III dataset (Alistair et al., 2016) and models specifically designed for it. The dearth of publicly accessible medical records explains this constraint. The medical notes in MIMIC-III only cover approximately 9,000 codes, about half of the full list of available ICD-9 codes.

Hence, there is a need for more ICD coding benchmark datasets, which will improve reproducibility, model comparisons, and inclusion of automated ICD coding in future studies.

MIMIC-III (Medical Information Mart for Intensive Care-III) is a collection of raw Electronic Health Records (EHR). Although the extensive adoption of EHR has resulted in the accumulation of vast amounts of data that can be used to develop predictive models to improve ICD coding, some data preprocessing must be carefully conducted for obtaining a benchmark set. For MIMIC-III, several benchmarks have been established for ICD-9 coding in full-code and high-frequent code settings (Mullenbach et al., 2018; Shi et al., 2017). These benchmarks have standardized the conversion of raw medical notes into data suitable for building predictive models. They offer clinicians and researchers easy access to high-quality data, accelerating research and validation efforts. Non-proprietary databases and open-source pipelines enable the reproduction and enhancement of clinical studies in previously unattainable ways. However, since ICD-9 was published in 1977, it includes outdated and obsolete terms. In contrast, ICD-10, launched in 1992, was designed to allow code expansion, allowing healthcare providers to employ codes more precisely tailored to patient diagnoses. Currently, the limited available benchmarks mainly focus on the coding settings of ICD-9, with no widely recognized benchmarks for ICD-10. A public ICD-10 benchmark would reduce entry barriers for new researchers and facilitate model development and comparison.

In this paper, we propose a public benchmark suite for ICD-10 coding using a large data set derived from MIMIC-IV (Johnson et al., 2023), the most recent public EHR data set containing a decade of critical care database. We standardize the pre-processing for generating multi-label instances, each of which contains a discharge summary and a set of ICD codes. In addition to ICD-10 coding, we use MIMIC-IV data to create a new ICD-9 benchmark with more data points and a greater number of ICD codes than MIMIC-III. We implement and compare several popular methods for the ICD coding prediction tasks. In particular, we carefully study the best practice in applying some state-of-the-art models for ICD prediction.

Our open source code [1] allows users to follow our data processing steps, generate benchmarks, and reproduce our experiments. This study equips future researchers with information, recommendations, and protocols for processing raw data and developing ICD coding models efficiently.

[1] https://anonymized

MIMIC Full	III	IV-ICD9	IV-ICD10
# of documents	52,726	209,359	122,317
Avg words per doc.	1,462	1,460	1,662
Avg ICD codes per doc.	13.9	13.4	16.1
Unique ICD codes	8,921	11,331	26,096

Table 1: Statistics of multi-label instances of MIMIC-III (ICD-9 only) and MIMIC-IV (both ICD-9 and ICD-10).

The paper is structured as follows: Section 2 introduces our pipeline for processing the raw MIMIC-IV data and provides relevant statistics. In Section 3, we present the baseline models used and showcases the benchmark results. Additionally, this section includes experiments on hyperparameter search, target-metric selection, and ablation studies. Section 4 highlights important related works. Finally, Section 5 concludes the paper.

2 ICD Code Benchmark

2.1 Data Processing

Unlike MIMIC-III, which exclusively contains ICD-9 codes, MIMIC-IV encompasses both ICD-9 and ICD-10 codes. The dataset includes 209,359 hospital admissions with ICD-9 codes and 122,317 hospital admissions with ICD-10 codes, compared to MIMIC-III's 52,726 documents. Additionally, there are seven admissions with both ICD code versions. In terms of labels, MIMIC-IV has 11,311 and 26,096 codes for ICD-9 and ICD-10 versions, respectively, while MIMIC-III has only 8,921 codes. Table 1 displays the basic statistics of the dataset, juxtaposing them with MIMIC-III's statistics. Overall, MIMIC-IV features more documents and labels for both ICD-9 and ICD-10 code versions, although the number of words and labels per data instance is roughly equivalent. Evaluating existing methods from MIMIC-III in the MIMIC-IV context is advantageous for determining their performance in larger and more complex multilabel classification scenarios.

We standardize the terminology as follows: patients are identified by their `subject_id`, and each patient may have multiple hospital admissions, denoted by `hadm_id`. Both `subject_id` and `hadm_id` can be traced back to the MIMIC-IV database to follow a patient throughout their hospitalization.

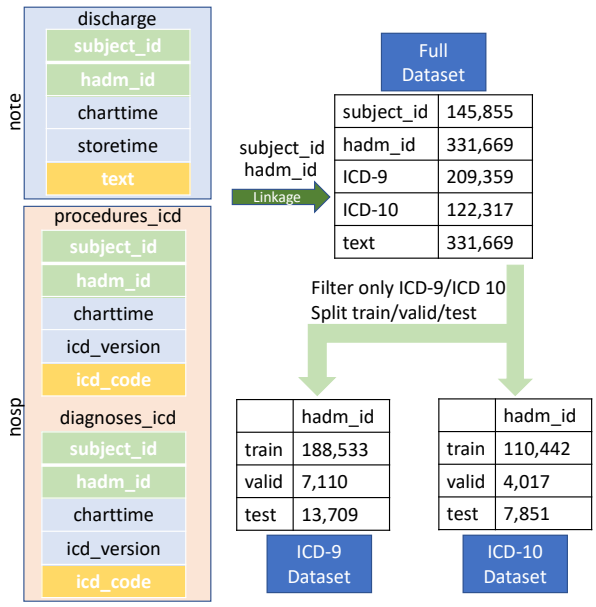

Figure 2: The workflow of data processing from raw data

The data processing workflow, illustrated in Figure 2, creates a data set consisting of discharge notes and ICD codes. Therefore, each multi-label instance corresponds to a discharge note and its respective ICD codes.

The construction is achieved by linking the 'discharge' table from the 'note' module to the 'procedures_icd' and 'diagnoses_icd' tables[2], if available, in the 'hosp' module, using `subject_id` and `hadm_id` as primary keys. Each discharge note corresponds to one pair of hospital admission id (`hadm_id`) and patient id (`subject_id`). By using the `hadm_id` and `subject_id` identifiers, the discharge note can be linked to its corresponding ICD code labels. After filtering out discharge notes which do not link to any 'procedures_icd' or 'diagnoses_icd' tables, the final dataset comprises 331,669 hospital admissions for 145,855 patients.

After obtaining all multi-label instances, it is necessary to divide the data set into the train, validation, and test sets. We follow the splitting procedures employed by Mullenbach et al. (2018). First, the master dataset is split to ensure no overlap of patient data (as identified by its unique `subject_id`) across the training, validation, and testing sets. Second, the dataset is partitioned based on patient percentage: 90%, 3.33%, and 6.67% for training, development, and testing, respectively. Statistics of the resulting sets are given in Table 2. Because we are not allowed to directly release these sets, we disclose our split by sharing the `hadm_id` for each partition to enable reproduction.

[2]ICD codes are divided into two systems: diagnosis codes and procedure codes. The diagnosis system covers diagnostic coding, while the procedure system consists of inpatient hospital procedure coding.

2.2 ICD Code Processing

In the previous section, we delved into full-label settings. Here, we generate datasets comprising only the top 50 most frequent labels, mirroring previous work with MIMIC-III. Specifically, we acquire the 50 most common codes from all instances, then refine the training, validation, and test sets to include instances with at least one of these codes. Table 2 presents the statistics of the resultant sets.

In addition to the top 50 datasets, we also prepare codes for studies utilizing code ontology (Vu et al., 2020) or descriptions for ICD code prediction (Yuan et al., 2022). Code IDs and descriptions were obtained from the latest ICD Code release by the Centers for Medicare and Medicaid Services[3][4]. The parental codes for the ICD-9 diagnosis codes were represented using the first four characters for codes starting with 'E', and the first three for the others. For example, diagnosis code E801.3 (Railway accident involving collision with other object and injuring pedal cyclist) belongs to the E801 category (Railway accident involving collision with other object), and diagnosis code 339.2 (Post-traumatic headache) has its parent code 339 (Other headache syndromes). For ICD-9 procedure codes, we use the first two characters as the parent codes, e.g., procedure code 08.01 (Incision of lid margin) belongs to category 08 (Operations on eyelids). For ICD-10, we use the first three characters as parent codes for both types of ICD codes; for instance, code Z00.01 (Encounter for general adult medical examination with abnormal findings) has its parent code as Z00 (Encounter for general examination without complaint, suspected, or reported diagnosis). Statistics of the resulting code hierarchy can be found in Table 2.

2.3 Corpus

Tables 1 and 2 show that the new ICD-9 and ICD-10 datasets contain more than thrice and twice the number of examples compared to MIMIC-III, respectively. The number of ICD codes observed in the dataset also increases. This is especially true for the transition to ICD-10, which features a richer and more specific hierarchy of diagnoses and procedures and thus more billable codes to select from.

[3]2023-icd-10

[4]latest-icd-9

	Train	Dev	Test	Train	Dev	Test
	MIMIC-III Full			**MIMIC-III 50**		
# Doc.	47,723	1,631	3,372	8,066	1,573	1,729
Avg. # of words per Doc.	1,434	1,724	1,731	1,478	1,739	1,731
Avg. # of parent codes per Doc.	13.7	15.4	15.9	5.3	5.6	5.7
Total # of unique parent codes	1,149	741	850	39	39	39
Avg. # of child codes per Doc.	15.7	18.0	17.4	5.7	5.9	6.0
Total # of unique child codes	8,692	3,012	4,085	50	50	50
	MIMIC-IV-ICD9-Full			**MIMIC-IV-ICD9-50**		
# Doc.	188,533	7,110	13,709	170,664	6,406	12,405
Avg. # of words per Doc.	1,459	1,472	1,460	1,499	1,516	1,501
Avg. # of parent codes per Doc.	12.1	12.2	12.0	4.6	4.7	4.7
Total # of unique parent codes	1,230	954	1,041	37	37	37
Avg. # of child codes per Doc.	13.4	13.5	13.3	4.7	4.8	4.8
Total # of unique child codes	11,145	5,115	6,264	50	50	50
	MIMIC-IV-ICD10-Full			**MIMIC-IV-ICD10-50**		
# Doc.	110,442	4,017	7,851	104,077	3,805	7,368
Total # of words per Doc.	1,662	1,671	1,642	1,687	1,695	1,669
Avg. # parent codes per Doc.	14.8	14.9	14.5	5.3	5.2	5.1
Total # of unique parent codes	2,220	1,449	1,627	38	38	38
Avg. # child codes per Doc.	16.1	16.2	15.8	5.4	5.4	5.3
Total # unique child codes	25,230	6,738	9,159	50	50	50

Table 2: Statistics of MIMIC-III and MIMIC-IV datasets under ICD-9 and ICD-10 settings.

Similar to MIMIC-III, these codes follow a natural long-tail distribution, where few codes appear often and the overwhelming majority is rare. Specifically, 50% of the ICD-10 codes appear in at most three discharge summaries (as compared to 12% for ICD-9). Furthermore, 2.0% ICD-9 and 6.3% ICD-10 codes appear only in the respective full test sets, which requires *zero-shot learning* approaches to correctly predict these.

Given the domain, the vocabulary of the discharge summaries is expectedly similar to MIMIC-III, with 42% and 40% of tokens[5] in MIMIC-IV-ICD9 and MIMIC-IV-ICD10 appearing in the vocabulary of MIMIC-III, respectively. Vocabulary differs less for frequent terms, with the overlap for the top 100 terms being 72% for ICD-9 and 64% for ICD-10 in MIMIC-III.

3 Empirical Study

3.1 Baseline Methods

In this section, we present the best existing models for ICD prediction, which are utilized for comparison in our study.

3.1.1 Models without External Data/Knowledge

The following models were trained without incorporating any form of external data or knowledge.

[5]after white-space tokenization, lower-casing stop-word and digit-only token removal

CAML The Convolutional Attention network for MultiLabel classification (CAML) was introduced by Mullenbach et al. (2018). It comprises a single-layer Convolutional Neural Network (CNN) and an attention layer that generates label-dependent representations for each ICD code.

LAAT The Label Attention Model proposed by Vu et al. (2020) consists of a single bidirectional Long Short-Term Memory (LSTM) network that produces latent representations for clinical notes. The label attention layer applies a structured self-attention mechanism to generate label-specific document representations.

JointLAAT The Hierarchical Joint Learning model of LAAT predicts the first level of ICD codes (parent labels) and uses them as additional input for the final label attention prediction. This approach helps address imbalanced and long-tail labels, training the model by minimizing the joint losses of both parent and child labels.

3.1.2 Models with External Data/Knowledge

The following models incorporate some form of external data or knowledge.

MSMN The Multiple Synonyms Matching Network (Yuan et al., 2022) leverages synonyms for improved code representation learning through a multi-head-synonym attention and pooling mechanism. The ICD Code synonyms required for the

Model	MIMIC-IV-ICD9-Full			MIMIC-IV-ICD9-50		
	F1		Precision	F1		Precision
	Macro	Micro	P@8	Macro	Micro	P@5
Models without External Data/Knowledge						
CAML (Mullenbach et al., 2018)	11.41	57.70	65.41	68.37	72.24	60.22
LAAT (Vu et al., 2020)	13.12	60.31	67.47	69.99	74.46	62.01
Joint LAAT (Vu et al., 2020)	14.17	60.37	67.46	69.93	74.33	61.95
Models with External Data/Knowledge						
MSMN (Yuan et al., 2022)	13.94	61.15	68.89	71.85	75.78	62.60
PLM-ICD (Huang et al., 2022)	14.40	62.45	70.34	71.35	75.46	62.44

Table 3: Results of ICD-9 code prediction models on the MIMIC-IV-ICD9-Full and MIMIC-IV-ICD9-50 test sets.

Model	MIMIC-IV-ICD10-Full			MIMIC-IV-ICD10-50		
	F1		Precision	F1		Precision
	Macro	Micro	P@8	Macro	Micro	P@5
Models without External Data/Knowledge						
CAML (Mullenbach et al., 2018)	4.61	53.32	65.44	65.13	69.80	62.08
LAAT (Vu et al., 2020)	4.47	55.40	66.97	68.15	72.56	64.39
Joint LAAT (Vu et al., 2020)	5.71	55.89	66.89	68.41	72.85	64.49
Models with External Data/Knowledge						
MSMN (Yuan et al., 2022)	5.42	55.91	67.66	70.31	74.15	65.16
PLM-ICD (Huang et al., 2022)	4.90	56.95	69.47	69.01	73.27	64.57

Table 4: Results of ICD-10 code prediction models on MIMIC-IV-ICD10-Full and MIMIC-IV-ICD10-50 test sets.

MSMN model (Yuan et al., 2022) are obtained from UMLS (Bodenreider, 2004).

PLM-ICD ICD Coding with Pretrained Language Models as proposed by Huang et al. (2022) is a framework that employs pretrained language models to encode documents and uses the label attention layer from Vu et al. (2020) to enhance ICD coding prediction.

3.2 Implementation and Evaluation

Details of data preprocessing steps and machines used for experiments are provided in the Appendix.

For evaluation metrics, we follow Mullenbach et al. (2018) to employ Macro- and Micro-F1, as well as Precision@k. Our results are based on multiple runs. We average the results from 5 runs.[6] Micro scores average the performance across all label-instance pairs, providing an aggregate measure of performance across all labels. On the other hand, macro-F1 calculates an unweighted average of F1 scores for each label, treating all labels as equally important. This is useful for datasets with imbalanced label distribution, which ensures that the performance on rare labels is not overshadowed by the performance on more frequent labels.

For Precision@k, we follow Mullenbach et al. (2018) to choose $k = 8$ for full settings and $k = 5$ for top-50 settings, which is based on the average number of labels per instance. Since MIMIC-IV and MIMIC-III have a similar number of labels per instance, we retain the same values of k. The stopping criterion is Precision@k on the validation set for CAML, MSMN, and PLM-ICD, and Micro-F1 for LAAT (based on original implementations).

3.3 Benchmark Results

We execute the baseline models using the original implementations with the hyper-parameters reported in their respective papers. The results are shown in Tables 3 and 4. Among the models without external data or knowledge, LAAT and Joint LAAT are superior to CAML in both the full and top-50 sets.

Among the models that utilize external data or knowledge, PLM-ICD achieves better performance in the full set, while MSMN shows better performance in the top-50 set. Also, except for Macro-F1, models that incorporate external knowledge tend to outperform those that do not.

[6] We do not consider AUC because the values are often too high to distinguish model performance.

Model	MIMIC-IV-ICD9-Full			MIMIC-IV-ICD9-50			MIMIC-IV-ICD10-Full			MIMIC-IV-ICD10-50		
	F1		Precision	F1		Precision	F1		Precision	F1		Precision
	Macro	Micro	P@8	Macro	Micro	P@5	Macro	Micro	P@8	Macro	Micro	P@5
CAML	12.19	58.67	67.37	69.08	73.82	61.48	5.17	54.60	67.46	66.64	71.57	63.32
LAAT	14.32	61.11	68.48	70.22	74.80	62.29	4.83	56.75	68.00	67.96	72.51	64.22

Table 5: Results of CAML and LAAT after parameter search on the MIMIC-IV test sets.

Validataion metric	MIMIC-IV-ICD9-Full			MIMIC-IV-ICD9-50			MIMIC-IV-ICD10-Full			MIMIC-IV-ICD10-50		
	F1		Precision	F1		Precision	F1		Precision	F1		Precision
	Macro	Micro	P@8	Macro	Micro	P@5	Macro	Micro	P@8	Macro	Micro	P@5
P@8	12.19	58.67	67.37	69.08	73.82	61.48	5.17	54.60	67.46	66.64	71.57	63.32
Micro-F1	13.08	58.95	67.03	68.95	73.78	61.39	5.05	54.51	67.43	66.81	71.63	63.40
Macro-F1	14.62	55.40	62.08	69.38	74.00	61.49	6.74	52.50	62.82	66.72	71.63	63.29

Table 6: Results of CAML using different validation metrics for parameter search on the MIMIC-IV test sets.

The superiority of LAAT over CAML on MIMIC-III was shown by Vu et al. (2020). Our results for MIMIC-IV support a similar pattern. However, in Section 3.4, we will demonstrate that the performance gap between the models becomes smaller once they are tuned.

3.4 Tuned Model Configurations

In Section 3.3, we directly run each baseline model. However, it is essential to conduct proper tuning for any machine learning method to assess the model's true capabilities. To establish a benchmark with reliable results for future development, we perform a comprehensive hyper-parameter search for CAML and LAAT. After identifying the configurations that yield the best validation results, we report the corresponding test performance. Note that the validation result of each hyper-parameter configuration is by predicting the development set shown in Table 2.

The results are presented in Table 5. Compared with Tables 3 and 4, where we apply the same hyper-parameters used in MIMIC-III, the performance consistently improves after a hyper-parameter search. For CAML, the improvement is significant, bringing it much closer in performance to LAAT. For LAAT, the improvement on the full set is larger than the top-50 set. In more difficult scenarios, like the full set with numerous labels, hyper-parameter search becomes crucial for better performance. In summary, we show the importance of hyper-parameter search in utilizing a model effectively, and our results can serve as a reference for future development.

In Section 3.2, we note that different validation metrics are used in the original implementation of baseline models. Ideally, the validation metric should correspond with the target metric (i.e., test metric in our case) for future prediction. For better understanding, we report CAML's test performance under various validation metrics in Table 6.

In general, the metric used in the validation process results in corresponding high test performance in the same metric, simply because the validation metric is optimized in the process. However, in some situations optimizing a metric such as Macro-F1 can result in a significant loss in other metrics. These experiments show that it is essential to decide the target measure according to the practical need. Also, when comparing different models, the same validation metric should be used.

3.5 Ablation study about code frequency

To assess the effectiveness of our baseline models, we perform an ablation study on both the MIMIC-IV-ICD9-Full and MIMIC-IV-ICD10-Full datasets, comparing the two best-performing models in our benchmark (MSMN and PLM-ICD). We examine the performance of these models on labels grouped by their frequency of appearance. To gain deeper insights into the models' predictions, we categorize medical codes into five groups based on their frequencies in MIMIC-IV-Full-ICD-9 and MIMIC-IV-Full-ICD-10 datasets: $1-10, 11-50, 51-100, 101-500, >500$. The

Frequency range	# ICD-9 codes	# ICD-10 codes
1-10	5,262	18,483
11-50	2,706	4,471
51-100	911	1,179
101-500	1,492	1,337
>500	853	626

Table 7: Label frequency distribution of MIMIC-IV-ICD9-Full and MIMIC-IV-ICD10-Full.

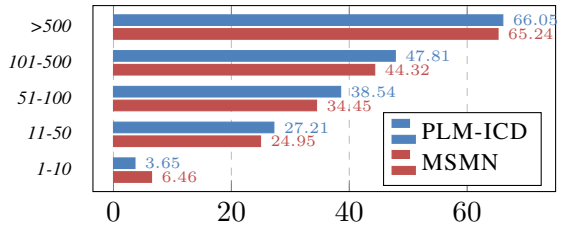

Figure 3: Comparison of Micro-F1 scores between PLM-ICD and MSMN on labels with different MIMIC-IV-ICD9-Full test set frequencies.

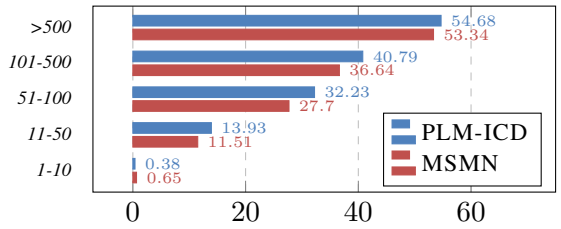

Figure 4: Comparison of Macro-F1 scores between PLM-ICD and MSMN on labels with different MIMIC-IV-ICD9-Full test set frequencies.

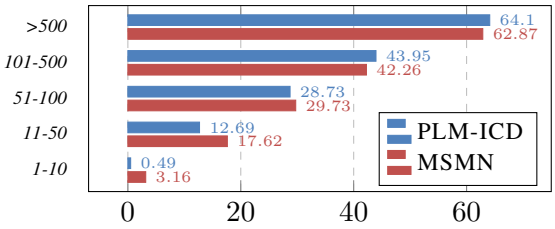

Figure 5: Comparison of Micro-F1 scores between PLM-ICD and MSMN on labels with different MIMIC-IV-ICD-10-Full test set frequencies.

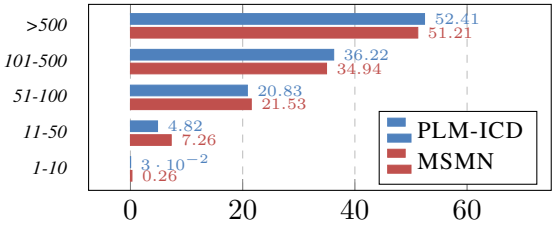

Figure 6: Comparison of Macro-F1 scores between PLM-ICD and MSMN on labels with different MIMIC-IV-ICD-10-Full test set frequencies.

statistics of all groups in both datasets are presented in Table 7.

MIMIC-IV-ICD-9 We compare the Micro-F1 and Macro-F1 scores across different groups. In general, PLM-ICD outperforms MSMN in most groups. For Micro-F1 shown in Figure 3, the differences are particularly noticeable in the frequent groups: 1% in the $>$ 500 group versus 3% in the $101-500$, $51-100$, and $11-50$ groups, while PLM-ICD performs worse than MSMN in the $1-10$ group, which contains the majority of codes. For Macro-F1 shown in Figure 4, the differences are similar in the frequent groups: 1% in the $>$ 500 group, 4% in the $101-500$ group, 5% in the $51-100$ group, and 2% in the $11-50$ group, while PLM-ICD still performs slightly worse than MSMN in the $1-10$ group. Overall, PLM-ICD learns better than MSMN in the ICD-9 setting.

MIMIC-IV-ICD-10 We compare the Micro- and Macro-F1 scores across different groups in Figures 5 and 6. In general, PLM-ICD outperforms MSMN in most groups. The differences in Micro-F1 are particularly noticeable in the more frequent groups (2% in $>$ 500 and $101-500$ groups versus), while PLM-ICD performs worse than MSMN in the $1-10$, $11-50$, and $>$ 500 groups, which contains the majority of codes. We observe the similar pattern in Macro-F1: the differences are 1%) in groups $>$ 500, $101-500$, while PLM-ICD performs worse than MSMN in the $51-100$ and $11-50$ group but slightly better in the $1-10$ group which contains the majority of codes. One possible explanation for this is that both models can learn from a few examples in very rare codes; however, with the assistance of multiple synonyms, MSMN can better match the semantic meaning of the codes to the medical notes compared to PLM-ICD, which does not consider code descriptions and relies solely on code embeddings. In the more frequent groups, PLM-ICD outperforms MSMN due to its superior encoder from large pretrained models. This suggests a potential future direction to improve both code representation and medical note representation using large language models.

4 Related Work

Before deep learning, automated ICD coding methods relied on rule-based or decision tree-based approaches (Farkas and Szarvas, 2008; Scheurwegs et al., 2017). The focus has since changed to neural networks, which can be classified into two main categories. The first involves encoding medical documents using pre-trained language models (Li and Yu, 2020; Liu et al., 2021), adapting pre-trained language models for the clinical domain (Lewis et al., 2020), or improving language models with medical knowledge, such as disease taxonomies, synonyms and abbreviations (Yang et al., 2022; Yuan et al., 2022). The second category aims to improve pre-trained language model representations by capturing the relevance between the document

and label metadata, including descriptions (Mullenbach et al., 2018; Vu et al., 2020), co-occurrences (Cao et al., 2020), hierarchies (Falis et al., 2019; Vu et al., 2020), or thesaurus knowledge such as synonyms (Yuan et al., 2022).

Many medical coding datasets exist in various languages and for various medical stages, but few of them are publicly available due to privacy concerns. The most popular datasets are the MIMIC databases. MIMIC-III was one of the first large, freely-available databases consisting of de-identified health-related data for patients admitted to critical care units at the Beth Israel Deaconess Medical Center from 2001 to 2012. The database includes information like demographics, vital sign measurements, laboratory results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of the hospital). MIMIC-III supports a wide range of analytic studies, including epidemiology, clinical decision-rule improvement, and electronic tool development. Mullenbach et al. (2018) and Shi et al. (2017) are the first two studies to publish a data pipeline for processing discharge summaries and matching them with ICD-9 codes, forming the MIMIC-III-full and MIMIC-III-top-50 sets, which became the popular benchmark for MIMIC-III ICD coding.

MIMIC-IV is the latest database containing real hospital stays for patients admitted to a tertiary academic medical center in Boston, MA, USA. It contains comprehensive information about each patient during their hospital stay, such as laboratory measurements, medications administered, and documented vital signs. The database aims to support a wide variety of research in healthcare. MIMIC-IV builds upon the success of MIMIC-III and incorporates numerous improvements. Several benchmarks and pipelines have been developed for MIMIC-IV to utilize its extensive dataset for various medical tasks: for example, Gupta et al. (2022) propose a data processing pipeline for extracting, cleaning, and preprocessing MIMIC-IV data for time-series tasks such as mortality prediction and readmission admission, while Xie et al. (2022) propose a benchmark for emergency department (ED) triage, critical outcome prediction, and reattendance prediction at ED triage. However, there is no benchmark for ICD coding for MIMIC-IV. Our work aims to provide a standard processing pipeline for this task, allowing researchers to process data, reproduce results, and conduct further research on top of it.

5 Conclusions and Recommendations

The field of machine learning is witnessing a surge in research focused on building clinical predictive models that effectively capture the complexities in EHR data and aid in predicting future outcomes. MIMIC datasets encourage research in this domain by providing a unique and extensive EHR dataset for researchers to explore. In this study, we establish a standardized benchmark for ICD coding on MIMIC-IV, covering both ICD-9 and ICD-10 codes. This process involves converting raw data into a task-specific format and applying popular deep learning baseline methods to the new datasets. Additionally, we demonstrate that code frequency not only emphasizes the model's enhanced predictive power for common codes but also suggests ways to improve performance for rarer ones. For example, MSMN performs better than PLM-ICD in predicting less common codes. Consequently, our benchmark dataset provides a more holistic and pragmatic approach to the ever-evolving labels in real-world applications. The long-tail distribution of ICD code predictions continues to challenge NLP, as traditional constraints might not fully address the breadth of real-world situations. Following the example set by Mullenbach et al. (2018), we make our data processing code open-source, enabling researchers to reproduce and enhance the results.

In the future, we plan to expand our benchmark by adding more baselines, potentially incorporating relevant features such as drug codes and patient vitals. Since ICD codes play a crucial role in enhancing patient care, facilitating research, and ensuring accurate communication among healthcare providers, our goal is to extend the use of clinical notes for joint prediction of ICD codes and readmission, triage, and mortality prediction tasks. By openly sharing our data processing code with the community, we hope to inspire others to join us in improving medical code prediction.

Limitations

Medical coding is crucial for the healthcare industry. With sufficient data for training and evaluation, automated medical coding can improve both accuracy and efficiency, aiding professional coders in reviewing patient medical records more effectively, reducing administrative costs, and ultimately improving care.

However, our study, driven by this objective, faces certain limitations that we address in the following. Like other data-driven studies, our result is constrained by our reliance on the settings of the MIMIC datasets. This dataset, characterized by its lack of diversity, includes only monolingual English discharge notes collected from emergency or intensive care units serving US patients. Consequently, it is challenging to assert with certainty that the effectiveness of state-of-the-art methods in this dataset would translate seamlessly into different clinical datasets, such as those encompassing other types of medical notes, languages, regions, or departments.

References

Johnson Alistair, EW, Pollard Tom, J, and al. et. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*.

O. Bodenreider. 2004. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, Online. Association for Computational Linguistics.

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

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. 2022. An extensive data processing pipeline for MIMIC-IV. In *Machine Learning for Health (ML4H)*, volume 193 of *Proceedings of Machine Learning Research*, pages 311–325.

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.

Alistair E. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and et al. 2023. Mimic-IV, a freely accessible electronic health record dataset. *Scientific Data*.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

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

Elyne Scheurwegs, Boris Cule, Kim Luyckx, Leon Luyten, and Walter Daelemans. 2017. Selecting relevant features from the electronic health record for clinical code prediction. *Journal of Biomedical Informatics*, 74:92–103.

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

Fu-Chiang Tsui, Michael M. Wagner, Virginia M. Dato, and Chung-Chou Ho Chang. 2002. Value of ICD-9-coded chief complaints for detection of epidemics. *Journal of the American Medical Informatics Association*, 9:S41–S47.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3335–3341.

WHO. 1948. International statistical classification of diseases and related health problems.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Ong, Fei Gao, and Nan Liu. 2022. Benchmarking emergency department triage prediction models with machine learning and large public electronic health records. In *American Medical Informatics Association Annual Symposium (AMIA)*.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

6 Appendix

Training Details For clinical note preprocessing, we employed the standard regular expression tokenizer from the Natural Language Toolkit (NLTK) to tokenize the text into a list of word characters, convert the text to lowercase, and truncate it to the maximum length for each model. For training, we primarily adjusted the batch size to accommodate our GPUs, as MIMIC-IV datasets are larger and contain more labels than MIMIC-III. CAML, LAAT, and JointLAAT were trained using a single 16GB Tesla P100 GPU. Meanwhile MSMN, unlike for MIMIC-III, required more than 32 GB of memory and was thus trained on an 80GB A100 GPU. PLM-ICD was optimized using two 16 GB V100 GPUs.

Parameter Algorithm and Search Space For the search algorithm, we employ grid search, selecting the parameter space by examining the most crucial hyper-parameters based on MIMIC-III-full.

The parameter tuning space of CAML is as follows:

Learning Rate	0.0001, 0.00001
Filter Size	8, 10, 12
Number of Filters	350, 550
Dropout	0.2, 0.6

For LAAT, the search space is as follows:

Learning Rate	0.001, 0.0003
Encoder Dropout	0, 0.2, 0.4
RNN Dimension	512, 768, 1024
d_a	256, 384, 512