Appendix

A Pseudo code of the DF-IAPF algorithm

We present the proposed DF-IAPF in Algorithm 1. In lines 7-9, this algorithm uses a span in clinical notes to obtain an n-gram as a phrase t and updates its occurrence in the local clinical note. Lines 10-12 update the global document frequency and phrase frequency. Finally, lines 15-16 calculate the DF-IAPF score for every phrase. In line 17 we sort all phrases descendingly by the DF-IAPF score and select the top phrases with the highest scores. Finally, we filter out shorter titles that are subsequences of longer titles with high scores in lines 18-20.

Note that this algorithm is an offline extraction of phrases before training. The computation procedures of DF-IAPF are similar to the TF-IDF, except that we add a for-loop of n-gram in line 3. Therefore, the time complexity of the DF-IAPF algorithm is $\mathcal{N} \times \mathcal{O}(\texttt{TF-IDF})$. In our experiments, the running time of the DF-IAPF algorithm is about 2 minutes.

Algorithm 1: Section Title Extraction

```
Input  : A set $\mathcal{S}$ of clinical notes $\mathcal{S} = \{S\}$;
         An integer $\mathcal{N}$ to control the maximum word count in n-grams;
         An integer $K$ to select top-$K$ phrases
Output : A candidate set $\mathcal{C}$ of section titles
1  NT $\leftarrow$ an empty mapping from phrases to counts with a default value of 0
2  APF $\leftarrow$ an empty mapping from phrases to a frequency list with a default value of an empty list
3  for $N \leftarrow 1$ to $\mathcal{N}$ do
4      for $S \in \mathcal{S}$ do
5          $n \leftarrow$ the number of words in $S$
6          PF $\leftarrow$ an empty mapping from phrases to frequencies with a default value of 0
7          for $i \leftarrow 1$ to $n - N + 1$ do
8              $t \leftarrow (w_i, w_{i+1}, \ldots, w_{i+N-1})$ // N-gram
9              PF($t$) $\leftarrow$ PF($t$) + 1 // Update the frequency of $t$ in this document $S$
10         for $t \in PF$ do
11             NT($t$) $\leftarrow$ NT($t$) + 1 // Update the frequency of documents containing $t$
12             Append TF($t$) to APF($t$) // Update the frequency list of $t$
13 $n_d \leftarrow |S|$
14 $\mathcal{C} \leftarrow$ an empty mapping from phrases to scores
15 for $t \in NT$ do
16     $\mathcal{C}(t) \leftarrow \frac{\mathrm{NT}^2(t)}{n_d \times \sum_{i=1}^{\mathrm{NT}(t)} \mathrm{APF}(t)_i}$ // DF-IAPF, Equation (1)
17 $\mathcal{C} \leftarrow$ Sort $\mathcal{C}$ descendingly by the score and select $K$ phrases with the highest scores
18 for $(t_1, t_2) \in \mathcal{C} \times \mathcal{C}$ do
19     if $t_1 \subsetneq t_2$ then
20         $\mathcal{C} \leftarrow \mathcal{C} \setminus \{t_1\}$ // Remove shorter titles that are subsequences of longer
            titles with high scores.
21 return $\mathcal{C}$
```

B Additional experiments

B.1 Dataset statistics

The detailed dataset statistics for each task are listed in Table 3.

B.2 Results of KEPT

We do not include KEPT [26] in the backbone models because our devices do not support the training of KEPT due to its high complexity. We list the result of KEPT (w/o CM) here for reference. It is worth

Table 3: Data statistics for the MIMIC-50, MIMIC-full, and MIMIC-rare-50 tasks.

Task	**Item**	**Train**	**Dev**	**Test**
MIMIC-50	# Doc.	8,066	1,573	1,729
	Avg. # words per Doc.	1,478	1,739	1,763
	Avg. # codes per Doc.	5.7	5.9	6.0
	Total # codes	50	50	50
MIMIC-rare-50	# Doc.	249	20	142
	Avg. # words per Doc.	1,770	1,930	2,071
	Avg. # codes per Doc.	1.0	1.0	1.0
	Total # codes	50	50	50
MIMIC-full	# Doc.	47,723	1,631	3,372
	Avg. # words per Doc.	1,434	1,724	1,731
	Avg. # codes per Doc.	15.7	18.0	17.4
	Total # codes	8,692	3,012	4,085

noting our proposed contrastive pre-training and masked section training are also applicable to KEPT.

- MIMIC-full prediction:
  - Macro F_1: 11.8
  - Micro F_1: 59.9
  - P@8: 77.1
  - P@15: 61.5
- MIMIC-50 prediction:
  - Macro F_1: 68.9
  - Micro F_1: 72.9
  - P@5: 67.3
- MIMIC-rare-50 prediction:
  - Macro F_1: 30.4
  - Micro F_1: 32.6

B.3 Extracted section titles

To demonstrate the effectiveness of our proposed DF-IAPF algorithm to extract section titles, we compare it with a rule-based extraction algorithm [23]. It designs special rules for every observed section title based on colons and occurrence frequencies to segment clinical notes into sections. We list the top 20 extracted section titles in Table 4.

Qualitative analysis Here, the rank is obtained using DF-IAPF scores (left) or occurrence frequencies (right). The symbol "+" indicates the title extracted by our DF-IAPF algorithm but not by the rule-based algorithm, while the symbol "−" means the title extracted by the rule-based algorithm but not the DF-IAPF algorithm in the top 20 section titles. In this table, we observe that 17 titles are commonly extracted by both algorithms, indicating that our automatic section title algorithm is comparable to the hand-crafted rule-based method in terms of effectiveness. We further analyze the rank of missing section titles from both algorithms in the top 20 titles. All the titles that are not extracted by DF-IAPF in the top 20 section titles appear in the top 30 titles. However, the titles that are missing in the rule-based method have very low ranks. It shows that even though the rules are carefully designed by humans, they may not be applicable to all clinical notes or titles. Therefore, we can conclude that our DF-IAPF algorithm is more universal than the rule-based method since it can effectively locate section titles and require less human effort.

Quantitative analysis To numerically demonstrate the effectiveness of our proposed DF-IAPF algorithm, we randomly select 50 clinical notes and manually extract the section title set Ω_i for each clinical note by medical experts. To evaluate the coverage of the top-20 extracted section titles $\hat{\Omega}$ by DF-IAPF and the rule-based method, we use an average intersection rate between Ω_i and $\hat{\Omega}$: $\frac{1}{50}\sum_{i=1}^{50}\frac{|\Omega_i|\cap|\hat{\Omega}|}{|\Omega|}$. The rate of DF-IAPF is 0.87, while the rate of the rule-based method is 0.83. The rates are less than 1 due to the absence of the bottom 3 titles in Table 4. Additionally, some clinical notes contain less frequent titles including "facilities", "addendum", etc. However, the rate of DF-IAPF is still higher than the rule-based method because "chief complaint", "discharge date", and "sex" are all top frequent section titles, while "discharge disposition" is a relatively less frequent title. Moreover, we report the frequency of section titles after segmentation using the 23 section titles in Table 4. We can see that all section titles have high frequencies. Together with the intersection

Table 4: Top 20 section titles extracted by our proposed DF-IAPF algorithm and a rule-based method using colons and occurrence frequencies.

Rank	DF-IAPF	Frequency	#	Rule-based	Frequency
1	history of present illness	0.95	1	admission date	1.00
2	date of birth	0.87	2	− *service*	0.95
3	+ *sex*	0.87	3	date of birth	0.87
4	+ *discharge date*	1.00	4	history of present illness	0.95
5	admission date	1.00	5	− *allergies*	0.87
6	social history	0.82	6	past medical history	0.90
7	past medical history	0.90	7	social history	0.82
8	discharge medications	0.83	8	− *discharge disposition*	0.75
9	medications on admission	0.77	9	discharge medications	0.83
10	discharge diagnosis	0.94	10	discharge diagnosis	0.94
11	discharge condition	0.85	11	medications on admission	0.77
12	discharge instructions	0.71	12	attending	0.71
13	major surgical or invasive procedure	0.78	13	family history	0.74
14	brief hospital course	0.98	14	discharge condition	0.85
15	pertinent results	0.68	15	discharge instructions	0.71
16	followup instructions	0.89	16	major surgical or invasive procedure	0.78
17	family history	0.74	17	physical exam	0.94
18	+ *chief complaint*	0.77	18	brief hospital course	0.98
19	attending	0.71	19	pertinent results	0.68
20	physical exam	0.94	20	followup instructions	0.89
23	service	0.95	38	chief complaint	0.77
28	discharge disposition	0.75	664	discharge date	1.00
29	allergies	0.87	1726	sex	1.00

rate, it further proves the coverage and accuracy of the extraction algorithm. Note that the rank of the section titles extracted by the rule-based method is different from the order of frequencies. It is because the rank is determined by the number of extracted section titles based on colons before segmentation. However, not all section titles are followed with a colon. Therefore, after segmentation, the frequencies may be different from title extraction.

It is worth noting that the top 20-30 titles mainly contain some special tokens, such as "[**first name3**]", which are masked tokens in the original dataset for privacy concerns. In the contrastive learning part, we do not use sections that have little relation to ICD codes, including "date of birth", "sex", "admission date", "discharge date", "attending" and "service", and use the remaining titles to pre-train the clinical note encoder. In the training of ICD coding models, we use all 23 section titles (top 20, 23, 28, and 29) so that we make the least change to the completeness of clinical notes. For some less frequent section titles such as "addendum" mentioned before, we do not segment sections by applying them as separators, but merge them with adjacent sections. In this way, the content of these sections is reserved for training.

B.4 Results of MIMIC-full prediction

We report the results of MIMIC-full in Table 5. Here, w/o CM and w/ CM mean the results without and with the proposed CM strategies, respectively. In this task, we directly use the w/o CM results from the MSMN paper [27]. For the w/ CM results, we report the result of one run since this experiment requires a lot of time. For the results of w/o CM, all the backbone models have a relatively low Macro F_1 score due to the large size of the label set and long tail distribution of ICD codes, while PLM-ICD is the best in terms of Micro F_1, P@8, and P@15. As for the result w/ CM, the cells with green color indicates an improvement. From the comparison, we notice the proposed contrastive pre-training and masked training can improve the performance of the backbone models, among which the Macro F_1 score is increased by 7.1% on average. However, the PLM-ICD model does not improve as much as other backbone models. We infer it is because the PLM-ICD model already split clinical notes into chunks with a fixed length. Even with our training strategies, it somewhat breaks

Table 5: Results (%) of MIMIC-full when trained with and without the proposed contrastive pre-training and masked training (CM) strategies. Cells with the green color denote an improvement of w/ CM compared to w/o CM. Here, we do not provide a p-value since we run backbone models one time.

Model	w/o CM				w/ CM			
	Macro F_1	Micro F_1	P@8	P@15	Macro F_1	Micro F_1	P@8	P@15
MultiResCNN	8.5	55.2	73.4	58.4	9.3	55.9	74.0	58.8
HyperCore	9.0	55.1	72.2	57.9	9.6	55.6	73.0	58.5
JointLAAT	10.7	57.5	73.5	59.0	11.5	58.3	73.9	59.4
EffectiveCAN	10.6	58.9	75.8	60.6	11.3	59.4	76.2	61.1
PLM-ICD	10.4	59.8	77.1	61.3	10.6	60.0	77.2	61.5
MSMN	10.3	58.4	75.2	59.9	11.4	58.8	75.6	60.2

the information between sections so that the variability cannot be largely reduced by our proposed training strategies.

C Broader Impacts

Ethical considerations While EHR data contains private information of patients, the MIMIC-III dataset used in this work as well as all backbone models is a publicly available dataset. It de-identified the sensitive information of patients and doctors with masks, including admission/discharge date, name, and hospital name (e.g., [**first name3**]) to protect privacy. Therefore, the data we used will not leak such information even if we publish our code and model parameters.

Societal Impacts Incorrect ICD coding can lead to medical billing errors which can affect patients and healthcare costs. However, as an enhancement of existing ICD coding models, our work aims to improve the prediction accuracy of ICD coding. We believe our method does not bring additional negative societal impacts to ICD coding.