NoteContrast: Contrastive Language-Diagnostic Pretraining for Medical Text

Prajwal Kailas* PKAILAS@BWH.HARVARD.EDU
Brigham and Women's Hospital
Max Homilius* MHOMILIUS@BWH.HARVARD.EDU
Brigham and Women's Hospital, Harvard Medical School
Rahul C. Deo RDEO@BWH.HARVARD.EDU
Brigham and Women's Hospital, Harvard Medical School
Calum A. MacRae CMACRAE@BWH.HARVARD.EDU
Brigham and Women's Hospital, Harvard Medical School

Abstract

Accurate diagnostic coding of medical notes is crucial for enhancing patient care, medical research, and error-free billing in healthcare organizations. Manual coding is a time-consuming task for providers, and diagnostic codes often exhibit low sensitivity and specificity, whereas the free text in medical notes can be a more precise description of a patient's status. Thus, accurate automated diagnostic coding of medical notes has become critical for a learning healthcare system. Recent developments in long-document transformer architectures have enabled attention-based deep-learning models to adjudicate medical notes. In addition, contrastive loss functions have been used to jointly pre-train large language and image models with noisy labels. To further improve the automated adjudication of medical notes, we developed an approach based on i) models for ICD-10 diagnostic code sequences using a large real-world data set, ii) large language models for medical notes, and iii) contrastive pre-training to build an integrated model of both ICD-10 diagnostic codes and corresponding medical text. We demonstrate that a contrastive approach for pre-training improves performance over prior state-of-the-art models for the MIMIC-III-50, MIMIC-III-rare50, and MIMIC-III-full diagnostic coding tasks.

Keywords: medical text, diagnostic coding, contrastive training.

1. Introduction

Accurate and automated diagnostic annotations of medical notes have become increasingly important in healthcare systems for improving the efficiency of care and enabling large-scale real-world data analyses. While diagnostic codes are vital for healthcare providers for tracking disease incidence and billing, manual coding is often cumbersome, time-consuming, and prone to errors. Machine learning methods have been developed to automate the diagnostic coding of medical notes. This is a complex task, as frequently multiple codes are applicable for a single note or single condition, and there are over 60,000 medical codes of varying specificity in the hierarchical International Classification of Diseases (ICD-10) system. While the large number of different codes allows for detailed and specific documentation of medical conditions, some codes are used distinctively by individual clinicians, or in different locations, and many are used infrequently. Often, only the most relevant codes for billing purposes are used per medical encounter. As a result, accurate coding can be seen as a multi-label problem with noisy training labels and a long tail of rarely applied diagnostic codes.

Recent work has used long-document transformers and contrastive pre-training to annotate notes spanning thousands of tokens. However, these approaches typically depend on pre-defined biomedical ontologies like the Unified Medical Language System (UMLS) (Yang et al., 2022; Yuan et al., 2022), or ICD-9 and ICD-10 hierarchies (Xie et al., 2019; Cao et al., 2020) to derive meaningful distances between different diagnoses, and often involve complex prepro-

* These authors contributed equally

cessing. We sought to combine a data-driven contextual embedding of diagnostic codes with a straightforward contrastive pre-training objective to improve the automated annotation of medical notes. For this we developed an approach based on i) contextual embedding models for diagnostic codes based on a large real-world data set ii) large language models for medical notes and iii) contrastive pre-training to build an integrated model of both ICD-10 codes and corresponding medical text. We show that this contrastive approach incorporating real-world data significantly improved performance over prior state-of-the-art approaches using static sources of biomedical information in the MIMIC-50, MIMIC-50 rare and MIMIC-full benchmarks.

2. Related work

2.1. Diagnostic Coding

Automatic diagnostic coding is a multi-label classification task assigning ICD codes to medical notes, typically ranging from several hundred to more than 2000 words per note. In addition, the label space is large and sparse, with over 60,000 codes in the most recent version of ICD-10, and a long tail of rarely diagnosed conditions. Several studies have proposed different methods to address this problem. Complex patterns between text and ICD codes were learned using variations of LSTM networks, dilated convolutions, residual connections, and per-label attention (Mullenbach et al., 2018; Li and Yu, 2020; Ji et al., 2020; Vu et al., 2020). Label representations were further improved by using graph convolution networks to capture the hierarchical structure of diagnostic codes (Xie et al., 2019; Cao et al., 2020; Michalopoulos et al., 2022). Shared representations can be further improved by extracting representations from low- and high-frequency codes via self-distillation (Zhou et al., 2021), and UMLS-based code synonyms have been used to provide more comprehensive knowledge than capturing code hierarchies (Yuan et al., 2022). Other studies have proposed various techniques to improve coding accuracy, such as using pre-trained biomedical language models with segment pooling to encode longer texts (Huang et al., 2022), exploiting the discourse structure of notes by utilizing section and reconciled label embeddings (Zhang et al., 2022), and incorporating tree-based features constructed from structured electronic health record (EHR) data such as lab values and medications as additional embedding vectors (Liu et al., 2022). Self-alignment learning with a hierarchical contrastive loss has been used to inject knowledge from biomedical ontologies, and prompt-based fine-tuning has been shown to be a powerful approach for predicting diagnostic codes (Yang et al., 2022).

2.2. Language Models for Biomedical Text

There are many pre-trained transformer-based language models for clinical and biomedical tasks, including those trained on masked language modeling (MLM) as the original BERT architecture (Devlin et al., 2019) and other pre-training objectives. Notable MLM models in the biomedical domain include SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019), and BioLM (Lewis et al., 2020), which are trained on the semantic scholar , scientific abstracts from PubMed and PMC or the MIMIC-III dataset (Johnson et al., 2016). Most of these models work at a sentence level and can handle up to 512 tokens, but for longer document-level tasks, the Clinical Longformer and Clinical BigBird (Li et al., 2022) models can handle up to 4096 tokens and are based on Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) architectures, respectively.

2.3. Medical Code Representations

Transformer models have also been used to encode temporal sequences of diagnostic codes across multiple hospital visits. These models are pre-trained with a masked language model objective on diagnostic, billing, and procedure code sequences. BEHRT (Li et al., 2020) is a BERT model that uses additional position and segment embeddings to distinguish between visits and encode temporal information. It was trained on EHR data from 1.6 million patients and tested on three disease prediction tasks (disease diagnosis on the next visit, within 6 months, and within 12 months). MedBERT (Rasmy et al., 2021) is conceptually similar to BEHRT but encodes both ICD-9 and ICD-10 codes and uses an additional serialization embedding to maintain the relative order of diagnostic codes within a visit. Lastly, G-BERT (Shang et al., 2019) combines graph neural networks and BERT to encode ICD-9 diagnostic codes and ATC drug classification codes. The model is evaluated on a multi-label medication recommendation task given medication and diagnostic history.

2.4. Contrastive Learning

Contrastive learning involves training a model on a batch of examples that contain positive and negative pairs, maximizing the agreement among positive pairs while minimizing the agreement between negative pairs. This method has been applied for general-purpose pre-training and in specific domains, such as medical imaging. The ConVIRT model (Zhang et al., 2020) was an early model to use a contrastive learning objective for medical visual representations, followed by several approaches focusing on medical images with accompanying textual descriptions such as radiology reports (Müller et al., 2021; Huang et al., 2021; Wang et al., 2021, 2022). CLIP (Radford et al., 2021) is a general large-scale contrastive learning model trained on 400 million image-text pairs from the internet and ALIGN (Jia et al., 2021) was trained on over one billion noisy image alt-text pairs. Contrastive approaches have also been applied to the image or text domain separately; for example, SimCLR (Chen et al., 2020) is a framework for visual representations based on image-image contrastive loss and augmentations. Text-text contrastive approaches sample pairs based on neighboring text segments on the internet (Neelakantan et al., 2022) or independent cropping data augmentations (Izacard et al., 2021) and have achieved state of the art results on classification, semantic search, sentence similarity, and retrieval. Contrastive learning is increasingly being applied to specific domains and shows promising results in improving performance on various tasks. For example, SCEHR (Zang and Wang, 2021) is a contrastive learning framework based on EHR time series data applied to clinical risk prediction problems.

3. Methods

We propose learning diagnostically relevant representations of medical notes by aligning medical text with corresponding sequences of one or more ICD-10 diagnostic codes used during the same clinical encounter. We use a contrastive learning approach for pre-training the model, where the associated ICD-10 diagnostic codes of a medical note are used as positive signal and contrasted against the diagnostic codes belonging to other medical notes. We describe this approach as a combination of three components, an encoder for sequences of ICD-10 diagnostic codes, an encoder for medical text, and a joint model for contrastive training and alignment of these components, each of which are described in more detail below.

3.1. Modeling ICD-10 Sequences

We trained a RoBERTa model (Liu et al., 2019) on temporal sequences of diagnostic codes using real-world data of a large patient cohort. To learn long-term temporal associations as well as co-occuring diagnoses, we used sequences of ICD-10 codes across multiple clinical encounters of a patient over time. We selected one encounter as the "encounter of interest", and calculated the time difference (in days) for past and future encounters, with 0 indicating all diagnostic codes in the current encounter. These relative position values were used to calculate positional embeddings based on sine and cosine functions of different frequencies (Vaswani et al., 2017). In addition, token type identifiers distinguished between the current encounter and others. This approach allowed us to encode past diagnostic history, future events, and patterns in the sequences. Figure 1A shows an example of the sequence of ICD-10 codes for a single patient, who had 6 hospital encounters in their history that have been coded, and the 4th encounter was randomly selected to be the current encounter. We consider each ICD-10 code as a single token, and trained the ICD-10 sequence model using the masked language modeling objective, where 20% of the ICD-10 codes in each sequence are masked out and the model needs to predict the original ICD-10 code of the masked token relying only on the surrounding context of codes. We increased the mask percentage from the standard 15% to 20% based on (Wettig et al., 2023) and evaluated the model using perplexity values during training. See Table 4 for hyperparameter choices for pre-training all models.

3.2. Modeling Medical Text

Since medical notes commonly contain more than 512 tokens, it was essential to develop models for medical text that can support much longer sequences (Figure 1B). We trained different models for medical text which support document lengths of up to 8192 tokens. We used the BioLM (Lewis et al., 2020) *RoBERTa-base-PM-M3-Voc-distill-align* as the starting model checkpoint. These models were pre-trained on text in PubMed, PMC, and MIMIC-III with a byte pair encoding vocabulary learned from PubMed. We converted them to a BigBird model to handle long sequences using the method presented in (Beltagy

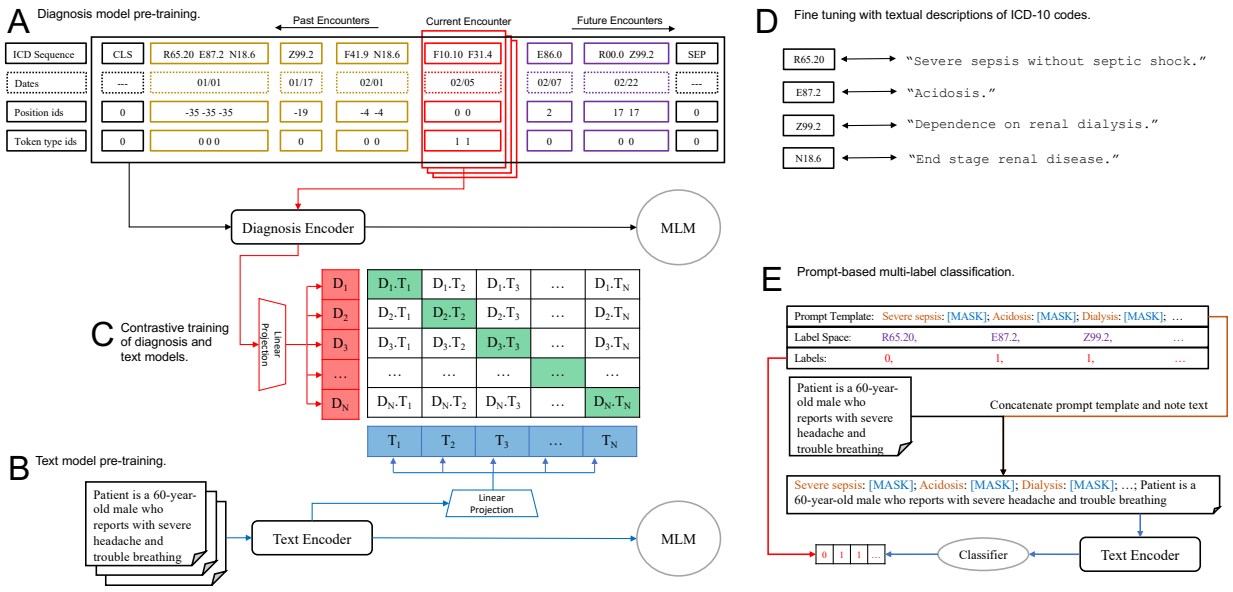

Figure 1: A) The input to the diagnosis model is a sequence of ICD-10 codes, where positions and token type ids are relative to a "current encounter" (shown in red). B) The text encoder is a language model for long documents pre-trained on medical notes. C) Contrastive training of corresponding ICD-10 code sequences and medical note pairs. D) Pairs of ICD-10 codes are matched to textual descriptions of the codes as a fine-tuning step. E) For prompt-based classification (Yang et al., 2022), the labels are concatenated with the medical note text. The prompt is a textual description of the ICD-9 code, and the outputs at the masked positions are processed to obtain the final multi-label classification output.

et al., 2020). In short, we repeatedly copied over 512-position embeddings and pre-trained the model on longer text using the MIMIC-III dataset before applying additional training objectives. We refer to this model trained solely on the MIMIC-III MLM task as *NoteLM*.

3.3. Contrastive Training

For the contrastive training, we used pairs of medical notes and their corresponding ICD-10 codes as positive pairs, and all other pairs as negatives. During training, we sampled batches of pairs and trained the model to predict which of the (text, ICD-10 code) pairs match across the batch (Figure 1C). This was done by jointly training the text and diagnostic encoders to maximize the cosine similarity between the text and code embeddings of the positive pairs while minimizing that of the incorrect pairs in the batch. We used the pre-trained transformer models described above to encode the text and ICD-10 diagnostic code sequence and used the hidden state of the CLS token to embed the ICD-10 sequence and the medical note. Compared to the diagnostic model pre-training, which used sequences spanning multiple medical encounters, we used only a single note and diagnostic codes from the same clinical encounter, not any past or future diagnostic codes, for the contrastive training step. The model representations are then projected into a multi-modal embedding space (T_N and D_N), and the InfoNCE loss (Oord et al., 2018) was computed as contrastive loss among positive and negative text-ICD and ICD-text pairs. In addition to the contrastive loss, we included the masked language modeling objective for the text model during training. This helped maintain existing textual properties while contrastively learning diagnostic properties. The contrastive and masked language losses were combined and weighted using un-

Table 1: Performance on MIMIC-III-50 dataset containing common ICD-9 codes. *NoteLM* and *NoteContrast* were run 5 times with different seeds to report mean and standard deviation. Performance of other methods is based on results collected from papers.

Model	AUC		F1		Precision
	Macro	Micro	Macro	Micro	P@5
JointLAAT (Vu et al., 2020)	92.5	94.6	66.1	71.6	67.1
MSMN (Yuan et al., 2022)	92.8	94.7	68.3	72.5	68.0
KEPT (Yang et al., 2022)	92.6	94.8	68.9	72.9	67.3
ISD (Zhou et al., 2021)	93.5 ±0.4	94.9 ±0.1	67.9 ±0.9	71.7 ±0.3	68.2 ±0.5
TreeMAN - all EHR (Liu et al., 2022)	93.7 ±0.2	95.3 ±0.0	69.0 ±0.2	72.9 ±0.2	68.2 ±0.1
TreeMAN - Text (Liu et al., 2022)	92.6	94.5	67.4	71.4	66.6
NoteLM 4k	92.3 ±0.10	94.3 ±0.07	65.2 ±0.35	71.1 ±0.22	66.7 ±0.11
NoteContrast 4k	93.1 ±0.09	94.9 ±0.06	67.4 ±0.28	72.6 ±0.24	67.6 ±0.17
NoteContrast 8k	93.5 ±0.14	95.3 ±0.06	68.7 ±0.46	73.4 ±0.18	68.1 ±0.21
NoteContrast 8k ICD	**93.8** ±0.04	**95.4** ±0.03	**69.2** ±0.21	**73.6** ±0.17	**68.6** ±0.18

Table 2: Performance on MIMIC-III-50-rare dataset containing uncommon ICD-9 codes. All methods were run 5 times with different seeds to report mean and standard deviation.

Model	AUC		F1		Initialization
	Macro	Micro	Macro	Micro	
MSMN (Yuan et al., 2022)	75.35 ± 1.32	77.41 ±0.66	15.3 ±2.77	16.65 ±1.48	Pre-trained
KEPT (Yang et al., 2022)	79.39 ±1.47	80.66 ±1.41	24.61 ±2.84	23.32 ±2.15	
NoteLM 4k	82.04 ±2.01	81.93 ±1.51	24.89 ±7.31	26.38 ±6.11	
NoteContrast 4k	**86.86** ±1.02	86.45 ±0.89	36.76 ±2.56	36.25 ±5.77	
NoteContrast 8k	85.65 ±1.16	**87.13** ±1.23	38.15 ±3.55	39.88 ±2.44	
NoteContrast 8k ICD	85.70 ±0.49	86.72 ±1.12	**39.08** ±2.15	**41.84** ±1.56	
MSMN (Yuan et al., 2022)	58.95 ±4.16	58.9 ±4.6	3.54 ±2.18	5.48 ±1.21	MIMIC-III-50
KEPT (Yang et al., 2022)	82.30 ±1.73	83.66 ±1.51	28.94 ±1.04	31.43 ±1.3	
NoteContrast 8k	88.40 ±0.33	**90.01** ±0.6	39.75 ±1.48	**43.3** ±1.5	
NoteContrast 8k ICD	**88.92** ±1.23	89.9 ±0.65	**40.26** ±2.96	42.64 ±1.97	

certainty weighting (Kendall et al., 2018). We trained three model versions to be able to compare performance on downstream tasks. We first trained the *NoteContrast 4k* model, which supports documents of length 4096. We started the training of this model using the NoteLM weights. We trained the model to minimize the weighted contrastive-ICD and MLM loss for 10,000 steps with a batch size of 64 and 16 gradient accumulation steps. Then, we converted the NoteContrast 4k model to handle sequences of length 8192 and trained the *NoteContrast 8k* model using the same loss for 10,000 steps with a batch size of 32 and 32 gradient accumulation steps. The *NoteContrast 8k ICD* model was based on the NoteContrast 8k model with additional fine-tuning using the contrastive loss objective. Here, the textual description for each ICD-10 diagnosis was treated as a very short medical note, and the associated ICD-10 code was considered a "sequence" of length 1 (Figure 1D). This step brought relevant codes/text closer and separated dissimilar codes/text, improving downstream task performance.

3.4. Prompt-based fine-tuning

We followed the prompt-based fine-tuning approach from Yang et al. (2022) for the ICD-9 coding task. Prompt-based fine-tuning is an alternative approach to multi-label classification where the multi-label classification task is reformulated as a cloze task. Each label is assigned a prompt template as shown in Figure 1E and the model fills the MASK token to indicate the presence or absence of the label in the multi-label space.

3.5. Pre-training Data

ICD-10 diagnostic codes. A RoBERTa-style model for ICD-10 sequences was trained on hospital encounters with ICD-10 codes, consisting of nearly 60 million real-world hospital encounters from 1.5 million patients. We first prepared a sequence for each patient containing all available ICD-10 diagnostic codes, leading to 1.5 million sequences of varying lengths. We randomly selected an encounter to be the "current encounter" and by changing the current encounter, generated 5 sequences per patient that contained the same sequence of codes, but different relative position and token type values, resulting in a final dataset of 7.5 million sequences.

Medical notes and contrastive training. The NoteLM and NoteContrast models were pre-trained on medical notes from the MIMIC-III dataset (Johnson et al., 2016), a collection of medical notes from over 40,000 patients. We used nearly 2 million notes for the MLM pre-training and around 50,000 notes for the contrastive language-diagnostic pre-training. Refer to Appendix A.2.2 for more details. For better clinical utility, we trained our diagnostic code model with a vocabulary of ICD-10 codes and translated ICD9 codes to ICD-10 codes[1]. To resolve ambiguous mappings, we selected an ICD10 code at random. We applied two pre-processing steps where we removed all de-identification placeholders and stripped extra white space. The NoteContrast 8k ICD model was also fine-tuned on textual descriptions of ICD-10 codes from the python package *icd10-cm v0.0.5*.

1. ICD-9 to ICD-10 mapping: https://github.com/AtlasCUMC/ICD-10-ICD9-codes-conversion/blob/master/ICD_9_10_d_v1.1.csv

4. Experiments

We conducted experiments with a series of NoteContrast models capable of handling document lengths of up to 4096 and 8192 tokens. Experimental results on the MIMIC-50, MIMIC-rare50, and MIMIC-III-full evaluations are shown in Tables 1, 2, 3. Our contrastive pre-training method demonstrated improved performance on most metrics in downstream ICD classification compared to the standard masked language modeling objective and existing state-of-the-art models for these tasks, including ISD, TreeMAN and KEPTLongformer (Zhou et al., 2021; Liu et al., 2022; Yang et al., 2022). It is important to note that our models were pre-trained on ICD-10 codes (i.e., have not seen ICD-9 codes during pre-training), and later fine-tuned for MIMIC-50, MIMIC-rare50 and MIMIC-III-full tasks with textual ICD-9 descriptions in the prompt based fine-tuning paradigm. Despite the ICD system change from pre-training to fine tuning, our model outperforms prior approaches developed specifically for ICD-9 coding. Implementation details are discussed in Appendix A.3 and Tables 5, 6 and 7 list hyperparameter choices selected based on dev set performance for the MIMIC-III tasks.

4.1. Dataset

We trained and evaluated our models using de-identified discharge summaries from the MIMIC-III dataset (Johnson et al., 2016), which has been widely adopted for benchmarking ICD-9 coding tasks. To allow comparison with other approaches, we adopted multiple tasks based on the MIMIC-III dataset: MIMIC-III-50, MIMIC-III-rare50, and MIMIC-III-full as previously described (Yang et al., 2022). The creation, size and preprocessing details of the datasets can be found in Appendix A.2.3

4.2. Metrics

We report micro and macro averaged F1 scores, micro and macro averaged AUC scores, precision at K (K = {5, 8, 15}), and recall at K (K = {8, 15}). All experiments were repeated 5 times with different random seeds (including model fine-tuning), and we present mean test results and standard deviation unless otherwise specified. The best thresholds for classification and computing precision, recall, and F1 were selected using the dev set for each task.

Table 3: Performance on MIMIC-III-full dataset, when using *NoteContrast* as a re-ranker of the top 300 *MSMN* predictions. Our model was run 5 times to report mean and standard deviation.

Model	F1		Precision	
	Macro	Micro	@8	@15
JointLAAT (Vu et al., 2020)	10.7	57.5	73.5	59
ISD (Zhou et al., 2021)	11.90 ±0.2	55.90 ±0.2	74.50 ±0.1	-
MSMN (Yuan et al., 2022)	10.3 ±0.3	58.2 ±0.4	74.9 ±0.3	59.5 ±0.1
PLM-ICD (Huang et al., 2022)	10.40 ±0.1	59.80 ±0.3	77.10 ±0.2	61.30 ±0.1
KEPT (Yang et al., 2022)	11.8 ±0.4	59.9 ±0.5	77.1 ±0.3	61.5 ±0.2
DiscNet+RE (Zhang et al., 2022)	**14**	58.8	76.5	61.4
NoteContrast 8k ICD	11.9 ±0.3	**60.7** ±0.03	**77.8** ±0.1	**62.2** ±0.1

4.3. Results

In the **MIMIC-III-50 task**, shown in Table 1, the *NoteContrast 8k ICD* model achieved a macro-AUC of 93.8 (+0.3), micro-AUC of 95.4 (+0.5), macro-F1 of 69.2 (+0.3), micro-F1 of 73.6 (+0.7), and precision@5 of 68.6 (+0.4). Numbers in parentheses show differences to prior best results. We excluded *TreeMAN - all EHR* from direct comparisons, since it uses text and structured data (e.g., lab results and medications), giving it additional information compared to the other methods. Under the **MIMIC-III-rare50** setting, shown in Table 2, the best performance was achieved by a *NoteContrast 8k ICD* model which had previously been fine-tuned on the MIMIC-III-50 task: macro-AUC of 88.92 (+6.62), micro-AUC of 89.90 (+6.24), macro-F1 of 40.26 (+11.32), micro-F1 of 42.64 (11.21). This approach performed better than fine-tuning the NoteContrast 8k ICD model on the MIMIC-III-rare50 dataset alone, which yielded macro-AUC of 85.70 (+6.31), micro-AUC of 86.72 (+6.06), macro-F1 of 39.08 (+14.47), micro-F1 of 41.84 (+18.52). For the **MIMIC-III-full** task, shown in Table 3, using the NoteContrast 8k ICD model to re-rank the top 300 candidate codes from MSMN improved over the previous state-of-the-art method for most metrics. We achieved macro-F1 of 11.9 (-2.1), micro-F1 of 60.7 (+0.8), precision@8 of 77.8 (+0.7), precision@15 of 62.2 (+0.7), recall@8 of 41.1 (+0.4), and recall@15 of 58.3 (+0.9). In agreement with prior work, we observed that prompt-based fine-tuning improved performance over traditional multi-label classification (Yang et al., 2022).

5. Discussion

Our final NoteContrast 8k ICD model combines a language model for clinical notes, contrastive training, and extends the document length to 8192 tokens. Below we compare different model versions to assess impact of each component.

Long-document language models offer a strong foundation for diagnostic coding of clinical notes. The NoteLM model, trained on 2 million notes, demonstrated comparable performance to earlier approaches such as ICD-BigBird, JointLAAT, and MSMN in the MIMIC-III-50 and MIMIC-III-rare50 tasks, without pre-training objectives related to diagnostic coding (Tables 1 and 2).

Diagnostic representation learning improved classification, especially for rare diagnoses. By incorporating a contrastive training objective between sequences of diagnostic codes and medical notes, we observed performance improvements across all evaluation tasks. Rare codes in particular (MIMIC-III-rare50, Table 2), had strong improvements compared to language model pre-training alone or prior work based on a hierarchical triplet-loss (e.g., KEPT). For example, contrastive training improved macro AUC by 4.82 points and macro F1 by 11.87 points for the NoteContrast 4k model compared to the NoteLM 4k model. Scaling the model from 4096 to 8192 tokens, combined with contrastive pre-training, further enhanced classification performance. The NoteContrast 8k ICD model with MIMIC-50 finetuning improved macro AUC by 6.88 and macro F1 by 15.37 points compared to the NoteLM 4k model.

Figure 2: Comparing outputs of the ICD-10 diagnosis model and text model before and after contrastive pre-training (left and right panels, respectively). We used a sample of 5000 ICD-10 codes as input to the ICD-10 encoder model, and textual descriptions of each code as input to the text model. For visualization, 768-dimensional outputs were projected using UMAP into two dimensions, scaled, and rotated using the Procrustes transformation. The contrastive pre-training step aligns corresponding output vectors in both models, leading to a more finely resolved embedding of all diagnoses.

Contrastive training aligns outputs from diagnostic code and text models. Joint training of text and diagnostic code models enabled alignment of their output vectors by maximizing the cosine similarity of positive pairs and minimizing that of incorrect pairs. Initially, the 2D UMAP projections of diagnostic code embeddings and their corresponding textual descriptions were dissimilar (Figure 2, left panels). While the text model formed clusters corresponding to high-level diagnostic categories, the UMAP embedding primarily separated injury-related codes (light blue) from other major diagnostic categories (Figure 2, bottom left). Conversely, the diagnostic code model assigned similar outputs to almost all non-injury codes and primarily separated different types of injuries (Figure 2, top left). However, after contrastive pre-training, embeddings of diagnostic codes and text were similar, resulting in overall better resolution when comparing diagnoses of different major categories (Figure 2, right panels).

5.1. Limitations

There are several limitations to consider in our study. Firstly, bias in the training data could affect the generalizability of our findings. The MIMIC-III dataset was sourced from a single medical institution and does not fully represent patient populations and healthcare practices found in other settings. Secondly, the long tail issue in the frequency of ICD

codes poses a challenge. Our model's representations for very rare ICD codes may not be accurate, as the training data did not contain many examples for these codes. Due to vocabulary cutoff during pre-training of the ICD10 diagnosis model, certain rare codes are likely missing from our diagnostic code model, limiting the coverage of our model. Another limitation is the relatively small batch size used for contrastive learning due to computational resource limits. While our model showed promising results with a batch size of up to 64 notes and diagnostic code sequences, other contrastive models have been pre-trained on much larger batch sizes and datasets (e.g., 32k pairs of images and text in CLIP). The limited batch size might impact the stability and convergence of the training process. Lastly, there are differences in the number and type of notes used for the MLM and contrastive training steps. While we used 2 million notes of multiple types, such as progress notes, radiology reports, and discharge summaries for MLM training, we used only 50,000 discharge summaries with corresponding ICD codes for contrastive training. This could introduce biases and limitations in our model when deployed on notes other than discharge summaries.

Using real-world clinical datasets, ideally incorporating multiple institutions from different geographic regions would yield much larger collections of medical notes and diagnostic codes, improving performance for rare diagnoses and better overall generalizability.

6. Conclusion

The InfoNCE objective has been related to maximization of mutual information between different modalities of the same concept (Oord et al., 2018), which is a powerful approach to utilize large collections of weakly labeled data. In this work, we built contextual embeddings of diagnoses based on their occurrence in a real-world data set, which has the potential to capture co-morbidities and temporally related diagnoses that occur in clinical settings. By aligning domain-specific text models with diagnostic code representations in a contrastive pre-training step, we were able to better annotate medical notes with diagnostic codes. The improvement was especially strong for infrequently occurring diagnoses, where our approach significantly improved over prior work that use fixed hierarchical sources of biomedical knowledge such as UMLS or ICD-9/ICD-10 ontologies. While this work focused on diagnostic coding of medical notes, data-driven contrastive pre-training can offer a powerful framework for other types of biomedical data with noisy labels.

Acknowledgments

This work was supported by One Brave Idea, co-founded by the American Heart Association and Verily with significant support from AstraZeneca and pillar support from Quest Diagnostics (to C.A.M. and R.C.D.). M.H. is supported by the Drs. Tobia and Morton Mower Science Innovation Fund Fellowship.

R.C.D. was supported by grants from the National Institutes of Health and the American Heart Association (One Brave Idea, Apple Heart and Movement Study) and is a co-founder of Atman Health. C.A.M. is supported by grants from the National Institutes of Health and the American Heart Association (One Brave Idea, Apple Heart and Movement Study), is a consultant for Bayer, Biosymetrics, Clarify Health, Dewpoint Therapeutics, Dinaqor, Dr. Evidence, Foresite Labs, Insmed, Pfizer and Platform Life Sciences, and is a co-founder of Atman Health. All other authors report no competing interests.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proc. of EMNLP*, pages 3615–3620, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv*, 2020. doi: 10.48550/arxiv.2004.05150.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proc. of ACL*, pages

3105–3114, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.282. URL https://aclanthology.org/2020.acl-main.282.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. of ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.clinicalnlp-1.2. URL https://aclanthology.org/2022.clinicalnlp-1.2.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3922–3931. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00391. URL https://doi.org/10.1109/ICCV48922.2021.00391.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv*, 2021. doi: 10.48550/arxiv.2112.09118.

Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 73–78, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.8. URL https://aclanthology.org/2020.clinicalnlp-1.8.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

Alistair E.W. 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. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:sdata201635, 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7482–7491. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00781. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *Proc. of ICLR*, 2015. URL http://arxiv.org/abs/1412.6980.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 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, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.17. URL https://aclanthology.org/2020.clinicalnlp-1.17.

Fei Li and Hong Yu. ICD coding from clinical text using multi-filter residual convolutional neural network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8180–8187. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6331.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1): 7155, 2020. doi: 10.1038/s41598-020-62922-y.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *arXiv*, 2022. doi: 10.48550/arxiv.2201.11838.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 2019. doi: 10.48550/arxiv.1907.11692.

Zichen Liu, Xuyuan Liu, Yanlong Wen, Guoqing Zhao, Fen Xia, and Xiaojie Yuan. TreeMAN: Tree-enhanced multimodal attention network for ICD coding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3054–3063, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.270.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of ICLR*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. ICDBigBird: A contextual embedding model for ICD code classification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 330–336, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.32. URL https://aclanthology.org/2022.bionlp-1.32.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proc. of NAACL-HLT*, pages 1101–1111, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint Learning of Localized Representations from Medical Images and Reports. *arXiv*, 2021. doi: 10.48550/arxiv.2112.02889.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and Code Embeddings by Contrastive Pre-Training. *arXiv*, 2022. doi: 10.48550/arxiv.2201.10005.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv*, 2018. doi: 10.48550/arxiv.1807.03748.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, 2021. doi: 10.1038/s41746-021-00455-y.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5953–5959. ijcai.org, 2019. doi: 10.24963/ijcai.2019/825. URL https://doi.org/10.24963/ijcai.2019/825.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for ICD coding from clinical text. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3335–3341. ijcai.org, 2020. doi: 10.24963/ijcai.2020/461. URL https://doi.org/10.24963/ijcai.2020/461.

Xiaosong Wang, Ziyue Xu, Leo Tam, Dong Yang, and Daguang Xu. Self-supervised Image-text Pre-training With Mixed Data In Chest X-rays. *arXiv*, 2021. doi: 10.48550/arxiv.2103.16022.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proc. of EMNLP*, pages 3876–3887, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.256.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.217.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP*, pages 38–45, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 649–658. ACM, 2019. doi: 10.1145/3357384.3357897. URL https://doi.org/10.1145/3357384.3357897.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. Knowledge injected prompt based fine-tuning for multi-label

few-shot ICD coding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1767–1781, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.127.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proc. of ACL*, pages 808–814, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.91. URL https://aclanthology.org/2022.acl-short.91.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

Chengxi Zang and Fei Wang. SCEHR: Supervised Contrastive Learning for Clinical Risk Prediction using Electronic Health Records. *2021 IEEE International Conference on Data Mining (ICDM)*, 00:857–866, 2021. ISSN 1550-4786. doi: 10.1109/icdm51629.2021.00097.

Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. 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, 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.254.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *Proceedings of the 7th Machine Learning for Healthcare Conference*, 182:2–25, 2020. URL https://proceedings.mlr.press/v182/zhang22a.html.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. Automatic ICD coding via interactive shared representation networks with self-distillation mechanism. In *Proc. of ACL*, pages 5948–5957, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.463. URL https://aclanthology.org/2021.acl-long.463.

Appendix A. Appendix

A.1. Ethics Approval

The study protocol was approved by the Mass General Brigham Institutional Review Board.

A.2. Dataset Processing

A.2.1. ICD-10 diagnostic codes.

ICD-10 codes from real-world hospital encounters from the MassGeneral Brigham hospital system were used to build the ICD-10 sequence dataset. We excluded any patient that had less than 5 hospital encounters. For each patient, we generated 5 sequences with the same set of codes but with different "current encounters" resulting in different relative position and token type values. The final dataset consisted of 7,502,320 sequences of hospital encounters from 1,500,464 patients of which 7,447,575 were used for training and the remaining 54,745 was used for validation. All sequences belonging to a given patient are either in the train or dev set (i.e., a patient cannot have sequences in both the train and dev set). Each ICD-10 sequence on average contained 86.64 ICD-10 codes.

A.2.2. Medical notes and contrastive training.

The MIMIC-III dataset (Johnson et al., 2016) contains 2,083,180 million de-identified notes. We removed all patients that appear in the test dataset of any evaluation task (MIMIC-50, MIMIC-rare50, and MIMIC-III-full). For masked language model (MLM) pre-training, the training set consisted of 2,059,772 notes and the dev set contained 20,036 notes. In the contrastive language diagnostic pre-training, we utilized only those notes that have diagnostic codes

available. We used a training set of 47,707 notes and a dev set of 1,631 notes. All notes that were used for contrastive pre-training were discharge summaries. The pre-processing of the data for both MLM and contrastive pre-training was minimal and only included removal of all de-identification placeholders present in the MIMIC dataset and stripping of extra white spaces.

A.2.3. Fine-tuning and evaluation

We used the steps described by (Yang et al., 2022) for creating the train, dev and test datasets for the MIMIC-50, MIMIC-rare50, and MIMIC-III-full tasks[2]. MIMIC-50 contains instances that had at least one of the top 50 most frequent codes. MIMIC-rare50 introduced by (Yang et al., 2022) is built by selecting the top 50 codes with less than 10 occurrences and contains instances that have at least one of these rare codes. MIMIC-III-full includes all discharge summaries. The processing of this dataset and the resulting train, dev and test splits have been used to benchmark multiple previous approaches (Mullenbach et al., 2018; Vu et al., 2020; Yuan et al., 2022).

A.3. Implementation Details

The ICD sequence encoder model was trained for 200K steps with a 2K batch size and took about 4 days of training time. The NoteLM and NoteContrast models were trained with a 1k batch size with varying gradient accumulation steps. NoteLM was trained for 7K steps and took about a day to train. The NoteLM model checkpoint with the lowest perplexity on the dev set was selected for all experiments. NoteContrast 4k and NoteContrast 8k were trained for 10K steps, NoteContrast 4k took about 2.5 days to train, while NoteContrast 8k took 5 days. NoteContrast 8k ICD was trained for 250 steps and took less than 5 minutes. The NoteContrast model checkpoints with the lowest contrastive loss on the dev set were used for all downstream evaluation tasks. We list the detailed hyperparameters for the pre-trained ICD and text models in Table 4. For the 3 downstream tasks (MIMIC-50, MIMIC-rare50, and MIMIC-III-full) we tuned the learning rate and weight decay using the dev set. The MIMIC-III-50 task took between 1-2 hours to complete training, the MIMIC-III-rare50 took between 25-50 minutes, and the MIMIC-III-full task took about 20 hours. The fine-tuning hyperparameters are listed in Tables 5, 6 and 7 for the MIMIC-III-50, MIMIC-III-rare50 and MIMIC-III-full tasks. The macro-AUC and micro-AUC performance on the dev set was used to select the model checkpoints for evaluation in the aforementioned tasks. All models were trained and fine-tuned on a DGX-2 node with 8 A100 GPUs. We used Adam (Kingma and Ba, 2015) as the optimizer with weight decay (Loshchilov and Hutter, 2019) for pre-training and fine-tuning all models. Our code is implemented based on PyTorch (Paszke et al., 2019) and Huggingface Transformers (Wolf et al., 2020) and available at https://github.com/obi-ml-public/NoteContrast. Since real-world data were used to train ICD-10 sequence models, we are unable to share the resulting model weights.

2. MIMIC-III preprocessing code: https://github.com/whaleloops/KEPT#download--preprocess-data

Hyper-parameter	ICD-10 Encoder	NoteLM	NoteContrast 4k	NoteC. 8k	NoteC. 8k ICD
Base Model	RoBERTa	BigBird BioLM	NoteLM	NoteC. 4k	NoteC. 8k
Dropout	0.1	0.1	0.1	0.1	0.1
Warmup Steps	4000	1000	1000	1000	50
Learning Rate	7.00E-04	5.00E-04	1.00E-04	7.50E-05	1.00E-04
Device Batch Size	64	64	64	32	1024
Gradient Accumulation	4	16	16	32	1
Effective Batch Size	2048	1024	1024	1024	1024
Weight Decay	0.01	0.01	0.1	0.1	0.1
Max Steps	200000	7000	10000	10000	250
Learning Rate Decay	Linear	Linear	Linear	Linear	Linear
Adam e	1.00E-06	1.00E-06	1.00E-06	1.00E-06	1.00E-06
Adam b1	0.9	0.9	0.9	0.9	0.9
Adam b2	0.999	0.999	0.999	0.999	0.999
Gradient Clipping	1	1	1	1	1
Maximum Sequence Length	512	4096	4096	8192	8192

Table 4: Hyperparameters for pre-training models. Base model represents the starting model checkpoint for pre-training. We convert the NoteContrast 4k model to support longer inputs (8192) before using it to train the NoteContrast 8k model.

Hyper-parameter	NoteLM	NoteContrast 4k	NoteContrast 8k	NoteContrast 8k ICD
Dropout			0.1	
Warmup Steps			200	
Learning Rate			2.50E-05	
Batch Size			64	
Weight Decay			0.01	
Max Steps			1500	
Learning Rate Decay			Linear	
Adam e			1.00E-06	
Adam b1			0.9	
Adam b2			0.999	
Gradient Clipping			1	
Maximum Sequence Length	4096	4096	8192	8192
Training Time	1 hour	1 hour	2 hours	2 hours

Table 5: Hyperparameters for fine-tuning NoteLM and NoteContrast models on MIMIC-III-50.

Hyper-parameter	NoteLM	NoteContrast 4k	NoteContrast 8k	NoteContrast 8k ICD
Dropout			0.1	
Warmup Steps			200	
Learning Rate			2.50E-05	
Batch Size			48	
Weight Decay			0.1	
Max Steps			500	
Learning Rate Decay			Linear	
Adam e			1.00E-06	
Adam b1			0.9	
Adam b2			0.999	
Gradient Clipping			1	
Maximum Sequence Length	4096	4096	8192	8192
Training Time	25 minutes	25 minutes	50 minutes	50 minutes

Table 6: Hyperparameters for fine-tuning NoteLM and NoteContrast models on MIMIC-III-rare50.

Hyper-parameter	NoteContrast 8k
Dropout	0.1
Warmup Steps	2000
Learning Rate	5.00E-05
Batch Size	192
Weight Decay	0.01
Max Steps	10000
Learning Rate Decay	Linear
Adam e	1.00E-06
Adam b1	0.9
Adam b2	0.999
Gradient Clipping	1
Maximum Sequence Length	8192
Training Time	20 hours

Table 7: Hyperparameters for fine-tuning NoteContrast 8k ICD on MIMIC-III-full.

Dataset	Train	Dev	Test
MIMIC-III-full	47,723	1,631	3,372
MIMIC-III-50	8,066	1,573	1,729
MIMIC-III-rare50	249	20	142

Table 8: Number of samples in each split of MIMIC-III-full, MIMIC-III-50 and MIMIC-III-rare50 datasets