Applications of BERT Models Towards Automation of Clinical Coding in Icelandic

Haraldur Orri Hauksson[1,2] **Hafsteinn Einarsson**[3]
[1] Department of Computer Science, ETH Zürich, Switzerland
[2] Department of Clinical Engineering and IT, Landspitali University Hospital, Iceland
[3] Department of Computer Science, University of Iceland, Iceland
haraldha@landspitali.is
hafsteinne@hi.is

Abstract

This study explores the potential of automating clinical coding in Icelandic, a language with limited digital resources, by leveraging over 25 years of electronic health records (EHR) from the Landspitali University Hospital. Traditionally a manual and error-prone task, clinical coding is essential for patient care, billing, and research. Our research delves into the effectiveness of Transformer-based models in automating this process. We investigate various model training strategies, including continued pretraining and model adaptation, under a constrained computational budget. Our findings reveal that the best-performing model achieves competitive results in both micro and macro F1 scores, with label attention contributing significantly to its success. The study also explores the possibility of training on unlabeled data. Our research provides valuable insights into the possibilities of using NLP for clinical coding in low-resource languages, demonstrating that small countries with unique languages and well-segmented healthcare records can achieve results comparable to those in higher-resourced languages.

1 Introduction

In recent decades, the healthcare industry's transition from paper-based to digital systems, particularly through the adoption of electronic health records (EHR), has opened up new avenues for accessing and utilizing healthcare data in research (Jha et al., 2009). This research has predominantly concentrated on structured data, including diagnostic codes and quantitative data from blood tests and other medical measurements. More recently, there has been a significant shift towards analyzing and using unstructured data, especially with the development of BERT-based models. During patient visits and treatments, medical staff compile clinical notes, consisting of unstructured, free-form text. This text often relates to diverse medical codes, like the International Classification of Diseases (ICD) diagnostic codes, encompassing over 70,000 entries.

Precise ICD coding is crucial for the healthcare industry. It plays a vital role in accurately recording patient medical histories, billing for treatments, and enabling research and analysis. Nevertheless, traditional clinical coding has been a manual, labor-intensive process prone to human error (Burns et al., 2012; O'malley et al., 2005; Cheng et al., 2009). Since the transition from paper to digital systems, automating clinical coding has been a key objective. The rapid advancements in Natural Language Processing (NLP), especially with the advent of Transformer-based models (Vaswani et al., 2017), have demonstrated increasing potential in meeting this automation goal for English (Huang et al., 2022).

In our study, we investigate this task for Icelandic, which is categorized as low to medium resourced language, by using over 10 million EHR notes spanning over 25 years from the Landspitali University Hospital (LUH). Previous studies have predominantly focused on English-language datasets, specifically MIMIC-III/IV (Johnson et al., 2016, 2023). For other languages, access to data can be a limiting factor and the datasets studied are usually not large. We review results established across a range of different languages in the literature review section. The primary objective of our study is to ascertain the performance attainable for the Icelandic language, through continued pretraining of existing models and various adaptations.

This challenge was approached with budgetary constraints in mind, due to the limited computational resources available for conducting on-site studies of sensitive EHR data. We explored various approaches, including continued pre-training on EHR data and conducting an ablation study to compare differences in the fine-tuning step. The highest-performing models attained a micro F1

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1956–1967

score of 72.2 and a macro F1 score of 65.2 for the top 100 labels and a micro F1 score of 66.2 and a macro F1 score of 18.3 for the top 8922 labels. We performed an ablation study which revealed that integrating LAbel ATtention (LAAT) led to enhanced classification performance (Vu et al., 2021). Particularly noteworthy is the fact that the performance mirrors results obtained from higher-resourced languages, underscoring the value of EHRs even for smaller countries with unique languages.

Due to the sensitive nature of the training data, the release of both the models and the data is not feasible. Nevertheless, we are confident that our findings provide valuable insights into the possibilities for languages with limited resources, such as Icelandic, in this field.

2 Related Work

2.1 BERT-based models for Icelandic

The development of BERT-based models for Icelandic has depended on access to large monolingual corpora. The largest manually curated corpus is the Icelandic Gigaword Corpus (IGC) compiled by the Árni Magnússon Institute that now encompasses 2.43 billion words (Steingrímsson et al., 2018; Barkarson et al., 2022). The IGC can be contrasted with the Icelandic Common Crawl Corpus (IC3) that was compiled from 63.5 million web pages belonging to the Icelandic top-level domain (.is). The IC3 was specifically created for studying the effect that different sources have on model pre-training (Snæbjarnarson et al., 2022).

Regarding transformer models, Snæbjarnarson et al. (2022) trained and released four monolingual RoBERTa models on these datasets providing baseline models for Icelandic. Continued pre-training of the multilingual XLMR-base model on IC3 was also studied with the resulting model showing comparative performance to the monolingual RoBERTa models.

2.2 Applications of language models in Healthcare

Language models have great potential to be applied on EHR data, primarily due to the performance of Transformer-based models, increased computational capacity, and the availability of extensive public datasets that have been used in the demonstration of state-of-the-art results (Johnson et al., 2016, 2023). For tasks such as ICD-code classification, the results are still far from perfect, especially on rare codes. However, they can create value in EHR interfaces, for example, they can be used to suggest ICD-codes when clinicians write a report.

Huang et al. (2022) put forth PLM-ICD, a framework for leveraging pre-trained language models (PLM) to tackle challenges encountered with automatic ICD classification. They show that using PLMs pretrained on domain-specific data provides performance improvement (an absolute increase of 5.7% and 1.5% respectively for micro and macro F1-scores) when compared to model that were pretrained on non-domain-specific data, but unfortunately, there's a lack of such models in low-resource languages and a lack of public clinical datasets, such as MIMIC-III/IV (Johnson et al., 2016, 2023) and n2c2 [1] being the biggest ones available.

Continued pretraining has turned out to be an improvement in other studies as well. Alsentzer et al. (2019) trained both BERT-base and BioBERT (Lee et al., 2020) on clinical EHR notes from the MIMIC-III dataset, getting the best performance from BioBERT, but showed that training BERT-base on out-of-domain clinical notes increased its performance across various tasks. In a similar vein, Lehman et al. (2023) found that smaller models pre-trained on clinical data outperform similarly sized general-domain models and show close performance to general-domain models of a much larger size. They also find that in-context learning of very large models such as GPT-3 for clinical tasks is not sufficient to replace fine-tuned clinical models. The issue on whether to pre-train on domain-specific data is not settled. To contrast with the results above, (Agrawal et al., 2022) have shown that general-domain Large Language Models (LLMs) perform well on zero-shot clinical tasks without specific training in that domain which raises the question if domain-specific clinical LLMs are needed.

Several studies have further explored pretraining on text from the medical domain. Zhang et al. (2020) introduced BERT-XML for automatic ICD classification, a model trained solely on over 7 million clinical EHR notes, tackling the domain problem that PLMs generally encounter. Furthermore, they improved performance by using the multi-label attention output layer from AttentionXML (You et al., 2019), initializing each label

[1] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

with the BERT embeddings of the text descriptions of the ICD codes. Similarly, Chirigati (2023) pre-trained a BERT-base architecture model on over 7 million notes from the same hospital and then fine-tuned it on various tasks, including readmission prediction where it greatly outperformed a group of physicians. Furthermore, they find that pre-training their model on a large corpus of data from multiple healthcare sites, and then fine-tuning models on local data for specific sites increases performance.

2.2.1 Applications in Other Languages

Table 1 gives an overview of studies that have been done on the medical code classification task in languages other than English. In the case of Icelandic, ICD classification has been the subject of two recent studies that are focused on a set of 4-6 ICD-10 codes in general practitioner (GP) EHR notes. Ellertsson et al. (2021) retrospectively compared the accuracy of GP's to ML models by hand annotating question-answer pairs to act as clinical features and then training an ensemble classifier which outperformed 6 physicians on the classification task. Furthermore, Hlynsson et al. (2022) expanded on this by using hand-annotated notes to train a Clinical Feature Extraction Model (CFEM) using IceBERT (Snæbjarnarson et al., 2022) to extract features on un-annotated notes and using logistic regression to label the notes with the predicted ICD-10 code.

Remmer et al. (2021) trained a classifier with the Swedish KB-BERT (Malmsten et al., 2020) using 6062 discharge summaries from gastrointestinal care units which had a zero F1-score on a label space of 263 values, but once aggregated to 10 code-blocks performed well. However, Tchouka et al. (2023), Coutinho and Martins (2022) and Suvirat et al. (2022) all trained ICD-10 classifiers using either monolingual or multilingual BERT models on corpus sizes of 56 to 169 thousand notes with various label counts and obtained macro F1-scores ranging from 38.5 to 88.2 depending on the size of the label space. Velichkov et al. (2020) compared different pre-trained language models for the task of ICD-10 classification from from diagnosis texts in Bulgarian, finding that the multilingual SlavicBERT (Arkhipov et al., 2019) which is trained on text in Bulgarian, Czech, Polish and Russian, was outperformed by MultilingualBERT (Devlin et al., 2019), BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019).

Other model architectures have also been applied in this area. Pribán et al. (2023) pitted a small Czech ELECTRA model against a Hierarchical Attention GRU model, with the latter overall outperforming the former, emphasizing the model size required to apply transformers. Reys et al. (2020) compared Logistic Regression, CNN, GRU, and CNN with attention (CNN-Att) on the task on Brazilian-Portuguese EHR notes, where the CNN-Att vastly outperformed the other with a micro F1-score of 48.5. Their method achieved a 1.3% absolute increase in performance over CAML, resulting in an F1-score of 53.7, when trained on the MIMIC-III dataset. Almagro et al. (2020) then took a different approach by extracting all sentences in Spanish EHR notes which included medical terms[2], then extracting features from them and applying more traditional methods such as SVMs and gradient boosting for classification.

3 Methods

3.1 Dataset

For this study, we use EHR data from Landspítali University Hospital in Iceland. The dataset includes all written records (16 million) dating back from 1997 to March 2023. The records in the dataset cover all aspects of hospitalization, which can be contrasted with public datasets such as MIMIC-III that focus on discharge summaries (Johnson et al., 2016)[3]. The records are composed of over 200 categories that have been developed and some even deprecated over the time period with the most common categories being Outpatient notes, Treatment notes, and Day-patient treatment notes. Around 40% of the records did not have any ICD code associated with them. The rest had one or more codes with a total of 10,520 unique codes used in the dataset.

The dataset was processed locally at the hospital on a single machine with an Nvidia RTX 4090 GPU. We used the following open source libraries in this work: Transformers and Datasets from HF, PyTorch, NumPy, Pandas and Scikit-learn.

[2]They used the IxaaMedTagger, a Spanish clinical part-of-speech tagging software - http://ixa2.si.ehu.eus/prosamed/resources

[3]Conventions can be different between healthcare institutions and in Iceland, clinical codes are usually not assigned to discharge summaries so we cannot make a direct comparison with MIMIC.

Language	Reference	Corpus size	Labels	Acc	Macro		
					Pre	Rec	F1
Br.-Portuguese	Reys et al. (2020)	69309	6918	-	-	-	-
Bulgarian	Velichkov et al. (2020)	345591	5854	81.9	-	-	86.0
French	Tchouka et al. (2023)	56014	6161	-	45.0	52.0	40.0
	-		1564	-	45.0	67.0	53.0
Icelandic	Ellertsson et al. (2021)	2563	4	-	-	-	-
	Hlynsson et al. (2022)	∼1200000	6	-	-	-	-
Portuguese	Coutinho and Martins (2022)	121536	1418	80.0	39.6	39.6	38.5
	-		611	83.7	53.1	49.9	50.1
	-		18	90.1	74.9	70.6	72.3
Spanish	Almagro et al. (2020)	7254	7078	-	69.5	-	-
Swedish	Remmer et al. (2021)	6062	263	-	0.0	0.0	0.0
	-		10	-	78.0	55.0	60.0
Czech	Pribán et al. (2023)	316808	1126	78.3	47.4	46.2	44.8
	-		1126	78.2	48.4	45.6	45.1
	-		1523	-	27.1	17.5	20.0
	-		1523	-	50.3	38.3	41.8
Thai	Suvirat et al. (2022)	91756	100	-	91.0	86.2	88.2
	-	148183	300	-	84.3	79.5	81.5
	-	168598	500	-	81.5	72.9	76.2

Table 1: An overview of recent studies applying NLP methods to the problem of medical code classification in non-English languages.

3.2 Data Pre-processing

The first step in the pre-processing pipeline was a cleaning step where occurrences of repeated characters that had been used to delimit sections of the text were removed[4]. HTML and XML segments were further removed and newlines were replaced by spaces.

After the cleaning step, we performed deduplication. We applied the MinHash approach which has been commonly used to deduplicate these types of datasets (Broder, 1997; Lee et al., 2022). We used the implementation in the `text-dedup` package (Mou et al., 2023). For the deduplication, we used a threshold of 0.5 for Jaccard similarity and the default amount of 256 permutations of hashing and an n-gram size of 5.

Due to the size of the dataset, it could not fit completely into memory on the machine (which had 64 GB of RAM). For that reason, we performed deduplication on subsets of the datasets that would fit in memory. We approached the deduplication both temporally and using random subsets. In the temporal step, we split the dataset into 8 parts with 25% overlap between contiguous parts. This step reduced the size of the dataset to 59.14% of its original size. After the temporal deduplication, we split the resulting dataset into 4 disjoint parts and performed deduplication on each one. We repeated that process four times and in total it reduced the dataset size to 59.04% of its original size. That is, random deduplication was not effective after performing the temporal deduplication step.

[4]This is a convention used by many professionals to structure the text.

3.3 Pre-training

Prior to fine-tuning, we performed continued pre-training on an existing BERT model to study its effect on downstream task performance. We split the dataset into 90% for training and 10% for validation. When continued pre-training was applied, the model was trained for 10 epochs in the standard masked language modeling task as in RoBERTa (Liu et al., 2019). The existing pre-trained model used in this study is the Icelandic BERT model `IceBERT` (Snæbjarnarson et al., 2022), and we refer to our further pre-trained model as `MedIceBERT` in this paper.

3.4 Classification

Despite language differences the dataset from LUH is considerably different in structure when compared to the MIMIC datasets, which in turn affects the underlying classification task. The MIMIC datasets do not represent EHR notes overall as it only includes critical care patients from a single hospital and prior research has been focused on classifying all ICD codes for whole admissions, with an average of 15 codes each. However, notes at LUH are shorter and have an average of 1.6 codes each.

For the classification task, we fine-tuned `MedIceBERT` on different dataset and label-set sizes to study how differences in these variables relate to classification performance. We used both the RoBERTa classifier[5] and the LAAT (Vu et al., 2021) classifier as implemented by Huang et al. (2022) (see details below). We used a 95/5 train/test split and we fine-tuned for ten epochs unless otherwise stated. During fine-tuning, we omitted notes with 20 or less words as well as truncated all tokens after 512, the context length of our RoBERTa models. This truncation affected 17% of our dataset.

To address the challenge of having a large label-space, we used the LAAT mechanism introduced by Vu et al. (2021) in our classifier. Prior work on ICD classification in English has shown that LAAT is an effective approach to learn label-specific features (Huang et al., 2022; Liu et al., 2022). We used the implementation by Huang et al. (2022), which is accessible online[6].

For a label space with ℓ labels, the label attention mechanism takes in $\mathbf{H} \in \mathbb{R}^{512\times768}$, the hidden representation of the last layer from our encoder, and computes the following

$$\mathbf{Z} = \tanh(\mathbf{H}\mathbf{W}^\top) \tag{1}$$

$$\mathbf{A} = \text{softmax}(\mathbf{Z}\mathbf{U}^\top) \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{d_a\times768}$ and $\mathbf{U} \in \mathbb{R}^{\ell\times d_a}$ are trainable parameters. We set the hyperparameter $d_a = 768$ as done by Huang et al. (2022). We proceed by computing

$$\mathbf{V} = \mathbf{A}^\top\mathbf{H} \tag{3}$$

where the i^{th} row represents the label-specific vector of the input document w.r.t. the i^{th} label. Finally, we output the vector

$$\left[\sum_{j=1}^{768}(\mathbf{L}\odot\mathbf{V})_{ij} + b_i\right]_{i=1}^{\ell} \tag{4}$$

where $\mathbf{L} \in \mathbb{R}^{\ell\times768}$ and $\mathbf{b} = [b_1, b_2, \ldots, b_\ell] \in \mathbb{R}^\ell$ correspond to trainable parameters. The i-th output of the result corresponds to the logit of the i-th label.

3.4.1 Top-100 Codes

We reduced the label scope by assigning all codes beyond the top 100 to a separate "rare disease" label and we also added a "no code assigned" label for notes with an empty label set. We trained `MedIceBERT` for 10 epochs on the classification task.

With an abundant amount of pre-training data, we studied the effect of continued pretraining before the fine-tuning phase by varying the size of the fine-tuning corpus from 32.5k to 1000k notes. For a standardized comparison, the training always had a fixed number of 1000k steps.

3.4.2 Full-set

To compare our models to those using the MIMIC-III full-set of ICD-10 codes, we fine-tuned our model using the top 8,922 labels in our dataset.

3.5 Evaluation

For measuring the model performance we report the following metrics: macro F1-score, micro F1-score, macro AUC[7] and micro AUC. We omitted precision@K which is commonly used for this task on the MIMIC-III dataset as 90% of our notes have 2 or fewer ICD codes and only 1% have 5 or more.

To facilitate comparison with prior results, we omit rare and no code labels when computing the performance metrics. This ensures that the performance metrics are not inflated due to potentially high performance on rare and no code labels. We report micro and macro F1-scores, as well as micro and macro AUC, for our `MedIceBERT` model on the top 100 and top 8922 label datasets.

Note that we do include rare and no code labels when we measure the effect of continued pretraining.

[5] https://github.com/huggingface/transformers/blob/main/src/transformers/models/roberta/modeling_roberta.py

[6] https://github.com/MiuLab/PLM-ICD/blob/master/src/modeling_bert.py#L96

[7] We used the method to calculate macro AUC by Mullenbach et al. (2018) which only averages the AUC score over those labels which were found in the dataset. https://github.com/jamesmullenbach/caml-mimic

4 Results

4.1 Classification Performance

The results of the 100-most-frequent and 8922-most-frequent models on their associated test sets are shown in Table 2 and Table 3. To understand how models in Icelandic compare to well-known results in English, we compare our results to models that were trained on the MIMIC-III dataset.

For our top 100 codes model, we achieve superior performance with respect to the AUC score when comparing to models trained on the MIMIC-III-50 dataset. For the F1-score, our model is close to the best performing models. This performance is noteworthy considering that our label-space is twice as large as those of the other models. Furthermore, when evaluating only on the top 50 labels, we see a modest improvement in the performance.

Similarly, our top 8922 codes model, which has the same label-space size as the MIMIC-III full dataset, achieved higher performance metrics compared to models trained on MIMIC-III. Our model outperforms the best-performing baseline by obtaining absolute increases of 6.4%, 7.9%, 0.9%, and 5.4% in micro-F1, macro-F1, micro-AUC, and macro-AUC.

4.2 Effect of Pre-training

We conducted an analysis to understand the impact of additional pretraining on the performance of models during fine-tuning. Specifically, we worked with 12 different models, each undergoing fine-tuning for 1 million steps. The duration of this fine-tuning varied between 1 and 32 epochs, depending on the size of the dataset used for fine-tuning. We adjusted the number of epochs for each model so that the total number of fine-tuning steps always summed up to 1 million. Six out of the twelve models corresponded to `MedIceBERT`, which was further pre-trained, and six corresponded to `IceBERT`, which was not pretrained explicitly on text from EHRs. As shown in Table 4, further pretraining improved the performance of our model compared to the base model across all dataset sizes. In this task, we fine-tuned the model to predict the presence of the top 100 ICD codes in a report, as well as to identify reports with no code or a rare code (i.e., a code not in the top 100). We note that repetitions of these experiments are required to get more robust results. The performance metrics for all models were calculated on the same evaluation dataset.

We note that `MedIceBERT` is exposed to text from the fine-tuning test set during the pre-training phase. However, it is not exposed to labels for the test set during that phase. An evaluation on notes that were written for a nine month period after the notes in the pre-training phase revealed a drop in performance. For 8922 codes, the F1-micro score decreased from 66.2 to 57.8; for 100 codes, it went from 72.2 to 66.5; and for 50 codes, it dropped from 74.9 to 69.9.

4.3 Ablation Study

To investigate the effects of using LAAT and training on unlabeled data, we conducted an ablation study on the `MedIceBERT` model, the results of which are shown in Table 5. These were performed on the top-50 (top 4 rows), top-100 (rows 5-8), and top-8922 (rows 9-12) label datasets. Note that in this ablation study, we fine-tuned the models for one epoch for the sake of time constraints.

We found that using LAAT overall improved our F1-scores and AUC metrics, with a smaller impact for smaller label-spaces but being crucial at 8922 labels. Furthermore we observed that training on unlabeled data reduced our F1-score across all label sizes, but instead increased our micro and macro AUC.

5 Discussion

This study has demonstrated the feasibility of automating clinical coding in Icelandic, a language with limited digital resources, by leveraging a substantial corpus of electronic health records. Our findings indicate that Transformer-based models, particularly when enhanced with label attention mechanisms, can achieve competitive results in clinical coding tasks. The best-performing model in our study attained micro and macro F1 scores that are not only competitive but also comparable to those achieved in higher-resourced languages. Specifically, when measuring performance using AUC, our models outperform the English state-of-the-art models. However, we note that due to structural and language differences, these direct comparisons need to be taken with a grain of salt.

The implications of our findings are significant for the field of NLP in clinical coding. They suggest that even languages with relatively small digital footprints can benefit from the advances in machine learning and NLP. This is particularly encouraging for smaller countries or those with unique languages, where digital resources may be scarce. Our

Model	Reference	Micro F1	Macro F1	Micro AUC	Macro AUC
LAAT	Vu et al. (2021)	71.5	66.6	94.6	92.5
JointLAAT	-	71.6	66.1	94.6	92.5
CAML	Mullenbach et al. (2018)	61.4	53.2	90.0	87.5
DR-CAML	-	63.3	57.6	91.6	88.4
Longformer	Ren et al. (2021)	68.9	63.1	93.1	90.5
BERT-ICD	Pascual et al. (2021)	-	-	88.7	84.5
HiLAT	Liu et al. (2022)	73.5	69.0	95.0	92.7
Ours (100)		72.2	65.2	99.5	99.3
Ours (50)		74.9	67.8	99.5	99.3

Table 2: Results on our 100-most-frequent dataset and MIMIC-III-50 test set. Note that *Ours (50)* still corresponds to the model trained with 100 labels, but the evaluation is only performed with respect to the top 50 labels.

Model	Reference	Micro F1	Macro F1	Micro AUC	Macro AUC
LAAT	Vu et al. (2021)	57.5	9.9	98.8	91.9
JointLAAT	-	57.5	10.7	98.8	92.1
CAML	Mullenbach et al. (2018)	53.9	8.8	98.6	89.5
DR-CAML	-	52.9	8.6	98.5	89.7
Longformer	Ren et al. (2021)	56.7	7.6	98.8	91.3
PLM-ICD	Huang et al. (2022)	59.8	10.4	98.9	92.6
Ours		66.2	18.3	99.8	98.0

Table 3: Performance on our 8922-most-frequent and MIMIC-III-full test set.

N	Epochs	IceBERT	MedIceBERT
31.25k	32	68.2	71.8
62.5k	16	69.3	74.0
125k	8	72.4	75.3
250k	4	74.4	77.3
500k	2	75.5	78.4
1000k	1	74.4	78.6

Table 4: Micro F1-Score for IceBERT and MedIceBERT over various fine-tuning regimes. The models were trained to predict the top 100 ICD codes, as well as rare and no code labels, resulting in a total of 102 classes. The dataset size (N) and number of epochs were varied while keeping the total number of training examples fixed at 1000k.

research demonstrates that with the right strategies, such as continued pretraining and model adaptation, it is possible to overcome the challenges posed by limited computational resources.

Our models, trained on a concise dataset of 100 labels, achieved accuracy comparable to those trained on the English MIMIC-III-50 dataset. This trend continued when expanding the label space to 8922 labels, mirroring the performance of models using the full MIMIC-III dataset. The success of our models can likely be attributed to the extensive dataset at our disposal and the brevity of Icelandic medical records, which may be more conducive to automated coding than the lengthy discharge summaries typically used. This has profound implications for the structure of healthcare data curation moving forward. In fact, recent methods have focused on segmenting clinical notes into sections to facilitate automated clinical coding (Lu et al., 2023). Our findings suggest that even languages with fewer resources can reach the performance of well-resourced languages if ample well-segmented data is available.

When examining how our work stacks up against

EC	LAAT	F1-Score		AUC	
		Micro	Macro	Micro	Macro
✗	✗	72.9	62.0	99.3	98.9
✓	✗	70.0	59.6	99.4	99.1
✗	✓	72.8	62.4	99.3	98.9
✓	✓	70.1	59.1	99.5	99.2
✗	✗	69.7	58.7	99.4	99.0
✓	✗	67.0	56.4	99.5	99.2
✗	✓	70.5	60.0	99.5	99.2
✓	✓	66.9	58.0	99.5	99.3
✗	✗	31.2	0.7	98.8	92.7
✓	✗	0.0	0.0	93.4	49.5
✗	✓	57.9	7.7	99.6	97.6
✓	✓	55.6	8.2	99.7	98.2

Table 5: Ablation study results on top 50, 100 and 8922 codes for fine-tuning `MedIceBERT`. EC stands for Extra Codes, i.e., training on notes with no assigned code or rare codes.

studies in other languages, we face the challenge of inconsistent performance metrics and tasks, as shown in Table 1. To facilitate more meaningful comparisons, we advocate for the adoption of uniform reporting standards in this research domain, such as consistently including both micro and macro F1 and AUC scores, akin to the convention in MIMIC dataset studies. While standardizing label space categories and training constraints could enhance comparability, we must balance this with the risk of imposing limitations that may not fit all research contexts.

Our ablation study focused on the impact of LAAT and the inclusion of unlabeled or out-of-scope data across datasets of varying sizes. We found that LAAT (Vu et al., 2021) had a marginal impact on smaller datasets but was essential for the expansive 8922-code dataset, likely due to the low frequency of certain codes. Intriguingly, training with out-of-scope data led to a slight decrease in F1-scores but an increase in AUC scores. An improvement in AUC could indicate that the model got better across all decision thresholds, but a decrease in F1-score could mean that it came at the cost of performance on low-frequency label categories. Further analysis will need to reveal the correct underlying nature of this tradeoff.

Our study confirms that additional pre-training on EHR notes using the masked-language objective markedly enhances code classification performance, aligning with prior findings on the benefits of domain-specific pre-training (Lehman et al., 2023; Huang et al., 2022; Yang et al., 2023; Kailas et al., 2023). We observed that our further-pretrained model consistently outperformed IceBERT by a margin of at least 3% in micro F1-score across all dataset sizes, from 31.25k to 1000k, after an equivalent training duration of 1000k steps. However, we also noticed a slight drop in the performance of `MedIceBERT` when applied to data newer than the pre-training data, suggesting the need for further research on the impact of distribution shifts in medical practice on model performance.

Our research was conducted under computational resource constraints, which limited the scope of our experiments and underscored the importance of powerful hardware in NLP research, particularly for large datasets. Nevertheless, in this resource-constrained scenario, we found temporal deduplication to be an effective data management strategy, which may be particularly suitable for medical datasets where note repetition is common.

Looking forward, potential research directions include aggregating all notes from a patient's stay to provide models with a richer context for code classification, contrasting with the current approach of analyzing individual notes. This could address the issue of incomplete code assignments in standalone notes. To manage the resulting long text sequences, techniques like document segment pooling (Huang et al., 2022), various text splitting strategies (Pascual et al., 2021), and alternative architectures capable of handling extended contexts, such as Mamba (Gu and Dao, 2023) and Hyena (Poli et al., 2023), could be explored.

In light of these findings and future research directions, we must also consider the ethical implications of automating clinical coding. As we move towards systems that can interpret and categorize medical text, questions of privacy, data security, and the potential for algorithmic bias must be addressed. Ensuring that these systems are transparent and equitable will be paramount, particularly as we extend these technologies to a broader range of languages and healthcare systems.

Moreover, the integration of NLP into clinical workflows raises important considerations about the role of human oversight. The balance between automated efficiency and expert judgment remains a delicate one and given the performance of our

models the current aim should be to augment human judgment rather than replace it.

6 Conclusions

In this paper, we presented the results for the first Icelandic models trained in the task of multilabel ICD-10 classification of EHR notes. We compared their performance on datasets with label spaces of 50-, 100- and 8922-codes, which were accumulated over 25 years at the Landspitali University Hospital. Our models performed similarly to state-of-the-art models trained on the MIMIC-III dataset. We explored how training on unlabeled data affected our model performance as well as how label attention impacted the confidence of our models in their choices.

We hope that our work provides a foundation and guidance for researchers working in other low- to medium-resource languages to explore the fine-tuning of pre-trained BERT-based models in their respective language on medical text for the task of clinical coding.

7 Acknowledgements

We would like to thank the Landspitali University Hospital for financing the hardware to make this study possible. We would also like to thank the Department of Clinical Engineering & IT at the Landspitali University Hospital for their assistance with data gathering for our research.

Limitations

Our study, while comprehensive, encounters several limitations that warrant discussion. Firstly, the quality and comprehensiveness of the EHR dataset pose significant limitations. The potential inaccuracies in labeling and representation within the EHR notes could impact the validity of our findings. More research will need to be done to uncover these issues, and the inherent limitations of real-world clinical data must be acknowledged.

The computational constraints we faced notably limited our exploration to base models only. This presents a considerable limitation, as larger or more complex models might yield drastically different results. All our experiments correspond to single runs and for that reason we could not provide error estimates for the performance metrics. The scalability of our approach to longer texts and larger datasets remains untested, marking a crucial area for future research.

A critical gap in our study is the determination of the practical performance threshold for clinical utility. It is not entirely clear what level of accuracy and reliability is needed for these models to be effectively implemented in a clinical setting. This gap highlights the need for ongoing collaboration with healthcare professionals to establish these benchmarks.

Moreover, our comparison of the Icelandic models with those trained on English datasets like MIMIC may not fully capture the nuances due to structural differences in datasets. This limitation in cross-language comparisons must be considered when interpreting our results.

We also acknowledge the potential for bias in our models, given the disparities that may exist in the dataset. These biases, if unchecked, could lead to skewed or unfair outcomes in clinical coding. While our Ethics Statement covers our commitment to addressing these concerns, they remain a pertinent limitation of our current study.

Lastly, the practical implementation of AI in clinical settings is fraught with ethical and operational challenges. Our study does not fully explore the implications of deploying these models in real-world settings, an area that requires thorough investigation and careful consideration.

While our study offers valuable insights into the application of Language Models in clinical coding for Icelandic, the limitations outlined above highlight the need for cautious interpretation and further research in this domain.

Ethics Statement

This research, conducted in adherence to the ACM code of ethics and professional conduct, strives to enhance the efficiency and accuracy of clinical coding in healthcare, ultimately serving the well-being of medical staff and patients. We acknowledge the sensitive nature of the EHRs used and have followed stringent data privacy and security measures to safeguard patient information. All data handling procedures were designed to comply with relevant legal and ethical standards in healthcare data management.

We recognize the potential for biases in the models, stemming from disparities in the dataset that reflect existing healthcare inequalities. These biases, related to demographics, socioeconomic status, and medical conditions, need to be critcally examined to mitigate their impact on the model's

fairness and accuracy before such a model is deployed. Our commitment to ethical AI extends to ongoing efforts to identify and rectify such biases, in collaboration with medical professionals who can provide invaluable insights and validation of our findings.

Transparency and reproducibility are key principles of our research, despite the inability to release the models and data due to privacy concerns. We have documented our methodologies and processes in detail, ensuring that other researchers can reproduce our study for other languages within ethical and legal boundaries.

The integration of AI in clinical coding bears significant implications for patient care, clinical practice and research on human diversity. Regarding deployment for clinical practice, we are acutely aware of the ethical responsibility this entails, including the potential risks and unintended consequences. We are working with healthcare experts to align our models with patient care priorities and clinical workflows.

Lastly, we emphasize the necessity for a cautious, measured approach to the deployment of such AI systems in clinical settings. Rigorous evaluation and validation, alongside ethical considerations such as non-discrimination and respect for patient privacy, are indispensable steps before these models can be considered for practical application. Through this research, we aim to contribute positively to the healthcare community.

Finally, we would like to state that GPT-4 was used as an aid in writing the code behind the experiments in this work, and to revise text in this manuscript.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mario Almagro, Raquel Martínez Unanue, Víctor Fresno, and Soto Montalvo. 2020. Icd-10 coding of spanish electronic discharge summaries: An extreme classification problem. *IEEE Access*, 8:100073–100083.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72.

Mikhail Arkhipov, Maria Trofimova, Yurii Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93.

Starkaður Barkarson, Steingrímsson Steinþór, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Finnur Ágúst Ingimundarson, and Árni Davíð Magnússon. 2022. Icelandic gigaword corpus (IGC-2022) - unannotated version. CLARIN-IS.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Elaine M Burns, E Rigby, R Mamidanna, A Bottle, P Aylin, P Ziprin, and OD Faiz. 2012. Systematic review of discharge coding accuracy. *Journal of public health*, 34(1):138–148.

Ping Cheng, Annette Gilchrist, Kerin M Robinson, and Lindsay Paul. 2009. The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. *Health Information Management Journal*, 38(1):35–46.

Fernando Chirigati. 2023. A language model for medical predictive tasks. *Nature Computational Science*, 3(7):576–576.

Isabel Coutinho and Bruno Martins. 2022. Transformer-based models for icd-10 coding of death certificates with portuguese text. *Journal of Biomedical Informatics*, 136:104232.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steindor Ellertsson, Hrafn Loftsson, and Emil L Sigurdsson. 2021. Artificial intelligence in the gps office: a retrospective study on diagnostic accuracy. *Scandinavian Journal of Primary Health Care*, 39(4):448–458.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hlynur Hlynsson, Steindór Ellertsson, Jón Daðason, Emil Sigurdsson, and Hrafn Loftsson. 2022. Semi-supervised automated clinical coding using international classification of diseases. In *Proceedings of*

the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022), pages 95–106.

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.

Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. 2009. Use of electronic health records in us hospitals. *New England Journal of Medicine*, 360(16):1628–1638.

A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark. 2023. Mimic-iv (version 2.2). PhysioNet.

A. Johnson, T. Pollard, and R. Mark. 2016. Mimic-iii clinical database (version 1.4). PhysioNet.

Prajwal Kailas, Max Homilius, Rahul C Deo, and Calum A MacRae. 2023. Notecontrast: Contrastive language-diagnostic pretraining for medical text. In *Machine Learning for Health (ML4H)*, pages 201–216. PMLR.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? *arXiv preprint arXiv:2302.08091*.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022. Hierarchical label-wise attention transformer model for explainable icd coding. *Journal of biomedical informatics*, 133:104161.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chang Lu, Chandan K Reddy, Ping Wang, and Yue Ning. 2023. Towards semi-structured automatic icd coding via tree-based contrastive learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Chenghao Mou, Chris Ha, Kenneth Enevoldsen, and Peiyuan Liu. 2023. Chenghaomou/text-dedup: Reference snapshot.

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.

Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639.

Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards bert-based automatic icd coding: Limitations and opportunities. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 54–63.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*.

Pavel Pribán, Josef Baloun, Jirí Martínek, Ladislav Lenc, Martin Prantl, and Pavel Král. 2023. Towards automatic medical report classification in czech. In *ICAART (3)*, pages 228–233.

Sonja Remmer, Anastasios Lamproudis, and Hercules Dalianis. 2021. Multi-label diagnosis classification of swedish discharge summaries–icd-10 code assignment using kb-bert. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1158–1166.

Weiming Ren, Tianshu Zhu, RUIJING Zeng, and Tongzi Wu. 2021. Towards transformer-based automated icd coding: Challenges pitfalls and solutions.

Arthur D Reys, Danilo Silva, Daniel Severo, Saulo Pedro, Marcia M de Sousa e Sá, and Guilherme AC Salgado. 2020. Predicting multiple icd-10 codes from brazilian-portuguese clinical notes. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 566–580. Springer.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir,

Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large icelandic text corpus. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Kerdkiat Suvirat, Detphop Tanasanchonnakul, Kanakorn Horsiritham, Chanon Kongkamol, Thammasin Ingviya, and Sitthichok Chaichulee. 2022. Automated diagnosis code assignment of thai free-text clinical notes. In *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE.

Y. Tchouka, J. Couchot, D. Laiymani, P. Selles, and A. Rahmani. 2023. Automatic icd-10 code association: A challenging task on french clinical texts. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 91–96, Los Alamitos, CA, USA. IEEE Computer Society.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Boris Velichkov, Simeon Gerginov, Panayot Panayotov, Sylvia Vassileva, Gerasim Velchev, Ivan Koychev, and Svetla Boytcheva. 2020. Automatic icd-10 codes association to diagnosis: Bulgarian case. In *CS-Bio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, pages 46–53.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3335–3341.

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

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 32.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: Large scale automated ICD coding using BERT pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 24–34, Online. Association for Computational Linguistics.