Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Ji, Shaoxiong; Hölttä, Matti; Marttinen, Pekka

Does the magic of BERT apply to medical code assignment? A quantitative study

Published in:
Computers in Biology and Medicine

DOI:
10.1016/j.compbiomed.2021.104998

Published: 01/12/2021

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Ji, S., Hölttä, M., & Marttinen, P. (2021). Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers in Biology and Medicine*, *139*, Article 104998.
https://doi.org/10.1016/j.compbiomed.2021.104998

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Computers in Biology and Medicine 139 (2021) 104998

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Does the magic of BERT apply to medical code assignment? A quantitative study

Shaoxiong Ji *, Matti Hölttä, Pekka Marttinen

Department of Computer Science, Aalto University, Espoo, 00076, Finland

ARTICLE INFO

Keywords:
Medical code assignment
Pretrained language models
BERT
Quantitative study

ABSTRACT

Unsupervised pretraining is an integral part of many natural language processing systems, and transfer learning with language models has achieved remarkable results in downstream tasks. In the clinical application of medical code assignment, diagnosis and procedure codes are inferred from lengthy clinical notes such as hospital discharge summaries. However, it is not clear if pretrained models are useful for medical code prediction without further architecture engineering. This paper conducts a comprehensive quantitative analysis of various contextualized language models' performances, pretrained in different domains, for medical code assignment from clinical notes. We propose a hierarchical fine-tuning architecture to capture interactions between distant words and adopt label-wise attention to exploit label information. Contrary to current trends, we demonstrate that a carefully trained classical CNN outperforms attention-based models on a MIMIC-III subset with frequent codes. Our empirical findings suggest directions for building robust medical code assignment models.

1. Introduction

Clinical notes generated by healthcare professionals are parts of electronic health records and provide an essential source for intelligent healthcare applications [1]. Medical information management aims to assign standard medical codes to each clinical document for categorization purposes, which requires professional medical knowledge and is usually costly and error-prone [2,3]. The International Classification of Diseases (ICD) system, as the most used coding system, provides a global standard for reporting diseases and health conditions. The rapid development of machine learning and natural language processing (NLP) can replace manual code assignment with automatic coding systems [4–6]. Practical medical code assignment requires to capture semantic concepts [7] and tackle the challenges of lengthy note encoding and large-dimensional code schemes.

Pretrained language models (PTM) such as BERT [8] learn contextualized text representation and have started a new era in NLP. NLP applications benefit from large-scale pretraining on massive corpora, and universal language representations from PTMs have been successfully utilized in downstream tasks via transfer learning. In the field of clinical NLP, incorporating pretrained contextualized language models to encode lengthy clinical notes for large-scale medical code prediction has not been well-studied. Recently, Refs. [9–11] performed preliminary experiments with pretrained models; however, these three pilot studies failed to achieve satisfactory results or provide in-depth analysis.

This paper investigates language models pretrained in various domains. A language domain here defines a distribution over a topical field such as biomedical documents and clinical notes. Specifically, we investigate the following three research questions (RQ1-3).

RQ1 *What kind of BERT pretraining works best?* We adopt domain-specific corpora and BERT variants pretrained with different domain adaptation illustrated in Fig. 1, and compare their performance on the MIMIC-III benchmark.
RQ2 *What kind of BERT fine-tuning formulation works best for long notes?* We employ classical fine-tuning, develop a hierarchical architecture for long clinical notes, and consider label-aware feature representation.
RQ3 *Are BERT models better than convolution-based approaches?* We reproduce convolutional neural networks (CNNs) with classical pretrained word embeddings and conduct a comparison.

Understanding medical text is a long-lasting research problem. We study an essential task in medical information management and diagnosis support - medical code assignment, assigning medical codes to clinical notes. Language models can either be pretrained using a large

---

* Corresponding author.
E-mail addresses: shaoxiong.ji@aalto.fi (S. Ji), matti.m.holtta@aalto.fi (M. Hölttä), pekka.marttinen@aalto.fi (P. Marttinen).

https://doi.org/10.1016/j.compbiomed.2021.104998
Received 24 September 2021; Received in revised form 26 October 2021; Accepted 26 October 2021
Available online 30 October 2021
0010-4825/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

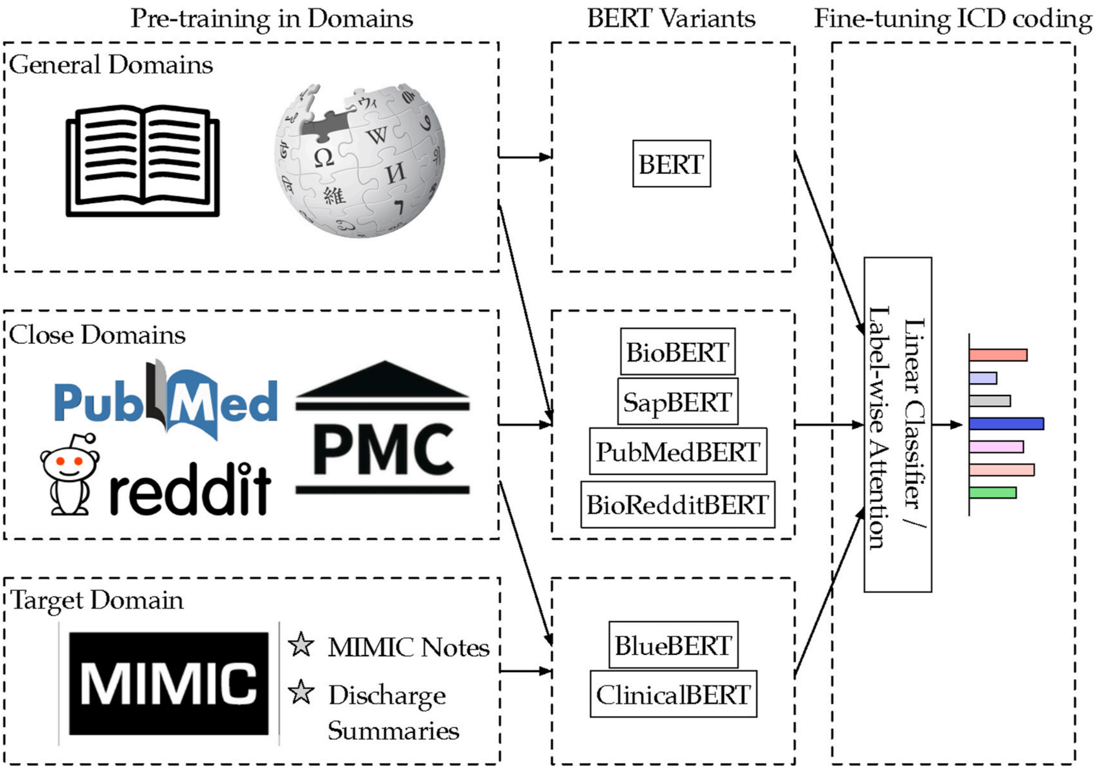


Fig. 1. Fine-tuning for ICD coding with pretraining in different domains: general, closely related, target, or mixed-domain pretraining.

corpus of medical text or trained from scratch and fine-tuned for a specific task. Several pretrained models have been published for medical NLP, and they differ in the collections of medical texts used for pretraining, e.g., biomedical, clinical, and medical-related social domains. We study the usefulness and relative merits of these different pretrained models and suggest improvements to the neural network architecture to improve performance with long notes: a hierarchical model for longer notes and label-wise attention to leverage relevant information about medical codes. Despite our careful attempts, we nevertheless find that fine-tuning pretrained models performs worse than carefully training conventional neural architectures from scratch. Hence, our results provide practical guidance for building medical information management systems: pretrained models offer a convenient plug-and-play solution; however, training robust existing standard models offers an appealing practical alternative with good performance in practice.

Our contributions are as follows.

- This paper conducts a comprehensive quantitative study to investigate the effect of knowledge transfer via mixed-domain and task-adaptive language model pretraining in different domains, and a thorough comparative study to answer the research questions.
- We propose a hierarchical BERT architecture with a label attention mechanism to enhance contextualized representation with label awareness for long clinical notes.
- We demonstrate that the classical CNN model with appropriate training can improve the predictive performance, achieving new state-of-the-art results on frequent medical codes (MIMIC-III top-50 dataset).

2. Related work

Rule-based and machine learning-based methods have been studied for diagnosis code assignment from clinical notes [12,13]. Ref. [13] proposed an SVM-based classification algorithm with a flat and hierarchy-based classifier. Recently, the research trend turns to deep neural networks. Convolutional neural networks are one popular category with many model architecture proposed, including CAML that applies CNNs and a label-wise attention mechanism [14], MultiResCNN that uses residual connection [9] and DCAN that utilizes dilated convolutions [15]. Recurrent neural networks are also extensively studied to capture sequential dependency in clinical notes. Such recurrent models include AttentiveLSTM [16], HA-GRU [17] and tree-of-sequences LSTM network [18] Attention mechanism for matching important diagnosis snippets is widely integrated into CNN- and RNN-based models [11,16]. CAML [14] introduced a label-wise attention mechanism to learn label-aware document representations.

Understanding clinical notes require professional medical knowledge. Many methods incorporate external knowledge sources to enhance neural architectures and facilitate clinical text understanding. Ref. [19] proposed a condensed memory network model with iterative condensation of external memory for network updating and data retrieval from Wikipedia. Ref. [20] used Wikipedia articles of medical codes to learn knowledge-aware embeddings jointly, and [21] utilized ICD code hierarchy with hyperbolic representation. Another direction to incorporate knowledge is through language pretraining and transfer learning. Refs. [9] and [10] reported preliminary results with semantic knowledge transferring. This paper conducts a comprehensive quantitative analysis.

Pretrained language models are trained on an auxiliary task, such as masked language modeling that predicts a word or sequence based on the surrounding context and gains improvement in many NLP tasks [22]. Pretraining such auxiliary tasks benefit from large-scale training on unlabeled corpora that are readily available from the web or textbooks. Ref. [23] hypothesized that pretraining acts as a type of regularization and found that pretrained models exhibit lower generalization errors on average. Several pretraining models in specific domains have been released, such as BioBERT [24] and ClinicalBERT [25]. They have also been applied in many domain applications; for example, [26] applied contextualized language models for phenotyping and [27] used pretrained models to encode clinical notes to predict hospital readmission.

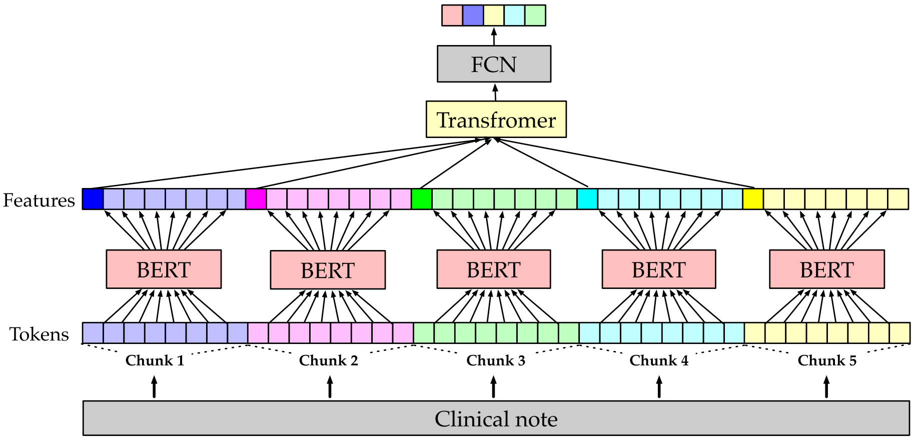


Fig. 2. Hierarchical BERT dealing with long clinical notes.

3. Method

We develop fine-tuning with different architectures, including a fully-connected classifier and a hierarchical classifier with an extra transformer atop (Fig. 2) to address the long-document challenge. Label-wise attention to learn label-aware document representations with these two fine-tuning architectures is described in Section 3.3.

3.1. Pretraining domains

We study three types of domains: 1) general domains such as book corpora and general Wikipedia articles; 2) domains that are closely related to the target clinical domain; 3) the target clinical domain. Assigning ICD codes from clinical notes is a task in the clinical domain. We consider biomedical and health-related social domains as candidate domains closely related to the clinical domain. Inspired by domain- and task-adaptive pretraining [28], we investigate different ways of pretraining models for medical code assignment: 1) pretrain only on general domains and immediately transfer to the target clinical domain; 2) continue pretraining on close domains and clinical domains such as the biomedical domain, and transfer to the target clinical domain 3) pretrain on close domains from scratch and transfer to the target clinical domain; 4) pretrain on mixed domains and further fine-tune on the target domain. The details of methods that fall into these classes can be summarized with three categories:

Pretraining in general domains does not involve specific topics or genres. We use BERT [8] pretrained on two unsupervised prediction tasks, i.e., masked language model and next sentence prediction, using the BooksCorpus [29] and English Wikipedia.

Mixed-domain pretraining contains a mixture of domains. We consider continued training and training from scratch. Domain-adaptive pretraining has been validated for the ability to improve the predictive performance [28] further. We use 1) BlueBERT [30] pretrained with PubMed text and MIMIC-III clinical notes; 2) BioBERT [24] continually pretrained on domain-specific data from PubMed abstracts and PMC full-text articles; 3) BioRedditBERT [31] initialized from BioBERT and continually pretrained on health-related posts from health-themed forums in Reddit; 4) PubMedBERT [32] domain-specific pretrained from scratch in biomedical domain using PubMed publications; and 5) SapBERT [33] that benefits from self-aligning biomedical entities to the Unified Medical Language System (UMLS) ontology and non-parametric metric learning, leading to a better-separated embedding space.

In-domain continued pretraining continues pretraining in the target clinical domain, which is also called task-specific pretraining [28, 34]. We use the ClinicalBERT [25] with the second phase of pretraining in the clinical domain using clinical notes and discharge summaries.

3.2. Fine-tuning with hierarchical structure

Clinical notes, consisting of patient history and discharge summaries, are often long documents. To address this, we develop two fine-tuning strategies, one based on the straightforward truncation and the other with hierarchical architecture. In the first approach, we truncate clinical notes to 512 tokens, take the final hidden state of the first token *[CLS]* as the pooled representation of the truncated note (denoted as $C \in \mathbb{R}^{d_h}$), and apply a fully connected network (FCN) as the classifier with sigmoid activation to predict output probabilities. This straightforward fine-tuning structure is denoted as *BERT-trun*, and it serves as the baseline.

Besides, we propose a hierarchical fine-tuning structure to deal with long notes, shown in Fig. 2. The lengthy clinical notes (with more than 512 tokens) are first divided into several shorter subsequences to build the lower-level contextualized representation. An additional transformer network [35] is built atop to capture the second-level sequential dependencies between the note segments. The classifier follows the same setup of the truncated version, i.e., FCN with sigmoid. We call this hierarchical structure *BERT-hier*.

The learning objective function adopts the binary cross entropy loss denoted as:

$$\mathcal{L} = \sum_{i=1}^{m} [- y_i \log(\widehat{y}_i) - (1 - y_i)\log(1 - \widehat{y}_i)], \tag{1}$$

where $y_i \in \{0, 1\}$ is the ground-truth label, $\widehat{y}_i$ is the sigmoid score for prediction, and m is the number of ICD codes. We use the AdamW optimizer [36] to fine-tune the model with backpropagation.

3.3. Label-wise attention

To further connect the document representation with label information, we introduce label-wise attention [14] in the fine-tuning procedures. The label attention network (LAN) prioritizes essential information in the hidden note representation relevant to the medical ICD codes. The LAN calculates the attention score $\mathbf{A} \in \mathbb{R}^{n\times m}$, which measures the importance between each pair of medical codes and words in the document. It is defined as a dot product:

$$\mathbf{A} = \mathit{Softmax}(\mathbf{HU}), \tag{2}$$

where $\mathbf{H} \in \mathbb{R}^{n\times h}$ is the hidden encoding of the BERT encoder's last layer, $\mathbf{U} \in \mathbb{R}^{h\times m}$ is the parameter matrix of the label attention layer (also known as the query), n is the number of tokens in the document, h is the hidden representation dimension, and m is the number of ICD codes. The attention layer's output is then calculated by multiplying attention $\mathbf{A}$ with the hidden representation from the last layer of BERT encoder. The attentive representation $\mathbf{V} \in \mathbb{R}^{m\times h}$ formalized as $\mathbf{V} = \mathbf{A}^{\mathrm{T}}\mathbf{H}$ is further used

Table 1
A statistical summary of datasets.

Dataset	Train	Dev.	Test	Labels
MIMIC-III top-50	8066	1573	1729	50
MIMIC-III full	47,723	1631	3372	8921

for medical code classification, representing sequential dependency and label awareness. The *BERT-trun* and *BERT-hier* fine-tuning architectures can both integrate the label-wise attention mechanism.

3.4. *Experimental setup*

This paper focuses on assigning ICD codes to textual discharge summaries from a hospital stay. The Medical Information Mart for Intensive Care III (MIMIC-III) repository [37] contains patients admitted to the Intensive Care Unit (ICU) at a US medical center from 2001 to 2012. The MIMIC database is available by request via this website https://mimic.physionet.org. We use the “noteevents” table in the latest release of version 1.4 with a total of 58,576 hospital admissions. Specifically, free-text discharge summaries are extracted. Two settings are used for experimental evaluation. One uses the top 50 frequent labels derived from Ref. [16]; while another uses the full set of ICD codes. Table 1 shows a statistical summary of two sets of MIMIC data.

3.4.1. *Preprocessing*

We use raw notes, ICD diagnoses, and procedures for patients from the public clinical MIMIC-III dataset for experiments. Discharge summaries labeled with a set of ICD-9 diagnosis and procedure codes include descriptions of procedures performed by a physician, diagnosis notes, patient’s medical history, and discharge instructions. Addenda of admissions are concatenated to a single document. The NLTK package is utilized for tokenization, and all tokens are converted into lowercase. Non-alphabetic characters, such as numbers and punctuations, are removed. All documents are truncated at 512 tokens and 2500 tokens for a single BERT encoder (*BERT-trun*) and hierarchical BERT encoders (*BERT-hier*).

3.4.2. *Training*

For the prior publications and the recommendation of fine-tuning the BERT model, we choose some common settings. For example, the dropout probability is 0.1. The Adam optimizer [38] is used to optimize CNN-based models, and the AdamW optimizer [36] to fine-tune BERT variants. We utilize a linear learning rate scheduler with warmup and layer-wise learning rates when fine-tuning BERT-based models. For retraining the CNN-based models, we use the CBOW of word2vec [39] and adopt static word embeddings. We set the batch size for MIMIC-III top-50 and full sets at 8, and the learning rate from $1e^{-6}$ to $1e^{-3}$.

All the models are run on a Linux cluster with Nvidia P100 or V100 GPUs. For the MIMIC-III top-50 data set, the fine-tuning architecture with truncated notes and a linear classifier has about 109 M parameters. In comparison, the MIMIC-III full set takes about 115 M parameters. For the hierarchical architecture, MIMIC-III top-50 and full codes data sets have 115 M and 128 M parameters, respectively. Generally speaking, fine-tuning pretrained models consumes the memory of a large GPU; however, it is arguably less expensive than training from scratch.

4. Results

We conduct a series of experiments with different pretrained models on various domains using two fine-tuning architectures and reproduce classical CNN-based models with word embeddings from scratch. We make our code publicly available at https://agit.ai/jsx/MCA_BERT.

Different evaluation metrics are utilized for experimental evaluation, including micro and macro F1 scores and area under the receiver operating characteristic curve (AUC-ROC). We evaluate the metrics of precision at k, where $k = 5$ for MIMIC-III subset with top-50 frequent codes and $k = 8$, 15 for full sets of MIMIC-III, given the observation that most medical documents are assigned no more than 20 codes. For example, the macro precision is calculated as the overall average precision across all labels, given by:

$$\text{Macro}-\text{Precision} \quad = \frac{1}{m}\sum_{\ell=1}^{m}\frac{\text{TP}_\ell}{\text{TP}_\ell + \text{FP}_\ell},$$

where TP_ℓ and FP_ℓ are the numbers of true and false positives of code ℓ. Micro scores give more weight to frequent labels by considering all labels jointly. For example, the micro precision is defined as:

$$\text{Micro}-\text{Precision} \quad = \frac{\sum_{\ell=1}^{m}\text{TP}_\ell}{\sum_{\ell=1}^{m}\text{TP}_\ell + \text{FP}_\ell}.$$

4.1. *Pretraining in close domains improves prediction (RQ1)*

This section studies pretraining in different domains (Sec. 3.1) to evaluate which pretraining scheme works best for medical coding in the clinical domain. Results on MIMIC-III top-50 and full code set are shown in Table 2, where hierarchical fine-tuning architecture is used. Overall, pretraining in mixed domains improves predictive performance to some extent over the BERT-base pretrained in general domains. These results show the effectiveness of transfer learning to enhance the learning capacity on downstream tasks. Among all pretrained models from various domains, PubMedBERT pretrained from scratch on biomedical article corpora gains a comparatively better performance. Specifically, it leads to improvements of 3.9% and 3.3% for F1 macro and micro scores on the MIMIC-III top-50 code set. One possible explanation is that specific domain pretraining makes downstream classifier concentrate on specified semantic knowledge. While for ClinicalBERT with three types of domain, the model’s attention may be distracted from relatively broad information. However, it still performs better than BERT-base with semantic knowledge only from general domains.

Table 2
Results of PLM fine-tuning with *BERT-hier* + LAN in various domains on MIMIC-III dataset with top-50 and full ICD codes. Clinical notes are truncated at length of 2500. Hierarchical fine-tuning structure with label-wise attention is used. **Bold** text denotes the best and *italic* text denotes the second best.

Model	MIMIC-III Top-50 Codes					MIMIC-III Full Codes					
	AUC-ROC		F1		P@5	AUC-ROC		F1		P@8	P@15
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro		
BERT-base	82.7	86.3	40.8	50.8	52.2	82.2	96.6	5.8	44.1	63.3	48.1
BlueBERT	**89.4**	**92.0**	61.0	65.6	62.8	84.4	97.5	5.1	42.5	62.6	47.3
BioBERT full text	88.8	*91.7*	60.4	66.0	*63.1*	85.2	97.4	**6.4**	**47.0**	*65.8*	*50.7*
BioRedditBERT	87.1	89.6	59.4	64.8	62.4	*86.5*	*98.0*	3.0	40.6	62.4	47.8
PubMedBERT full text	88.6	90.8	**63.3**	**68.1**	**64.4**	**87.4**	**98.1**	4.3	44.5	65.2	50.4
SapBERT full text	88.5	90.8	*62.2*	*66.7*	*63.1*	86.4	97.7	*6.2*	*46.8*	**68.5**	**53.0**
ClinicalBERT all notes	*89.2*	91.6	59.5	64.8	62.0	84.7	97.4	6.0	46.6	65.1	49.9

Table 3
Effect of more pretraining data with *BERT-hier* + LAN on MIMIC-III dataset with top-50 and full ICD codes. Clinical notes are truncated at length of 2500. **Bold** text denotes better performance.

Model	MIMIC-III Top-50 Codes					MIMIC-III Full Codes					
	AUC-ROC		F1		P@5	AUC-ROC		F1		P@8	P@15
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro		
PubMedBERT abstract	**90.1**	**92.5**	62.3	67.0	63.8	85.2	97.1	**5.8**	37.5	**66.6**	50.4
PubMedBERT full text	88.6	90.8	**63.3**	**68.1**	**64.0**	**87.4**	**98.1**	4.3	**44.5**	65.2	**51.7**
BioBERT abstract	**89.4**	91.6	59.5	64.8	62.7	84.5	97.3	6.3	46.5	65.3	50.0
BioBERT full text	88.8	**91.7**	**60.4**	**66.0**	**63.1**	**85.2**	**97.4**	**6.4**	**47.0**	**65.8**	**50.7**
ClinicalBERT disch. sum.	88.7	91.4	**60.6**	**65.9**	**62.6**	82.9	96.6	**6.2**	45.0	63.8	48.7
ClinicalBERT all notes	**89.2**	**91.6**	59.5	64.8	62.0	**84.7**	**97.4**	6.0	**46.6**	**65.1**	**49.9**

Table 4
Effect of more pretraining data with *BERT-trun* on MIMIC-III top-50 code set. Clinical notes truncated at length of 512. **Bold** text denotes better performance.

Model	AUC-ROC		F1		P@5
	Macro	Micro	Macro	Micro	
PubMedBERT abstract	**82.4**	**85.9**	47.5	54.8	54.7
PubMedBERT full text	82.1	84.4	**52.6**	**57.3**	**55.7**
BioBERT abstract	80.9	83.6	48.9	54.8	54.0
BioBERT full text	**81.8**	**84.3**	**50.5**	**55.4**	**54.5**
ClinicalBERT disch. sum.	81.8	84.2	49.7	55.8	54.9
ClinicalBERT all notes	**82.3**	**85.3**	**50.6**	**56.9**	**55.7**

Table 5
Results on MIMIC-III top-50 code set fine-tuning with *BERT-trun* and *BERT-hier* with FCN or LAN.

Model	AUC-ROC		F1		P@5
	Macro	Micro	Macro	Micro	
BERT-base-trun + FCN	80.1	83.0	46.3	51.6	51.1
BERT-base-hier + FCN	85.4	88.4	53.1	59.4	57.6
BERT-base-hier + LAN	83.7	86.5	48.1	54.2	53.5
BlueBERT-trun + FCN	78.2	82.1	23.7	36.6	45.9
BlueBERT-hier + FCN	89.1	91.8	59.9	65.1	62.5
BlueBERT-hier + LAN	**89.4**	**92.0**	61.0	65.6	62.8
BioBERT-trun + FCN	81.8	84.3	50.5	55.4	54.5
BioBERT-hier + FCN	87.3	89.5	59.9	65.6	62.7
BioBERT-hier + LAN	88.8	91.7	60.4	66.0	63.1
BioRedditBERT-turn + FCN	80.7	83.5	49.3	55.5	53.8
BioRedditBERT-hier + FCN	87.0	89.4	59.1	64.2	61.9
BioRedditBERT-hier + LAN	87.1	89.6	59.4	64.8	62.4
PubMedBERT-trun + FCN	82.1	84.4	52.6	57.3	55.7
PubMedBERT-hier + FCN	88.8	91.4	57.7	64.2	62.2
PubMedBERT-hier + LAN	88.6	90.8	**63.3**	**68.1**	**64.4**
SapBERT-trun + FCN	82.3	84.8	51.2	56.8	55.2
SapBERT-hier + FCN	88.5	90.3	**63.3**	67.3	63.8
SapBERT-hier + LAN	88.5	90.8	62.2	66.7	63.1
ClinicalBERT-trun + FCN	82.3	85.3	50.6	56.9	55.7
ClinicalBERT-hier + FCN	87.2	90.7	54.6	62.5	60.7
ClinicalBERT-hier + LAN	89.2	91.6	59.5	64.8	62.0

4.2. *Effect of pretraining with more data (RQ1)*

Can a larger pool of unsupervised pretraining data lead to performance gain in downstream clinical prediction? To answer this question, we conduct experiments on the effect of more pretraining data using three groups of pretrained model. In the biomedical domain, there are 200K abstracts from PubMed and 270K full-text articles from PubMed-Central. For the clinical domain, The CATEGORY value of the MIMIC-III dataset includes 'Discharge Summary', 'ECG', 'Radiology', and 'Echo'. 'Discharge summary' indicates that the note is a discharge summary. Besides, there are also other free-text notes. For example, 'Report' indicates a full report, and 'Addendum' indicates an additional text added to the previous report.

The results with *BERT-hier* fine-tuning architecture on two sets of the MIMIC-III dataset are shown in Table 3. We also study the effect using truncated short notes with fully connected tuning architecture, with results reported in Table 4. These two tables suggest that pretraining with more data leads to better performance in most cases.

4.3. *Hierarchical fine-tuning improves prediction (RQ2)*

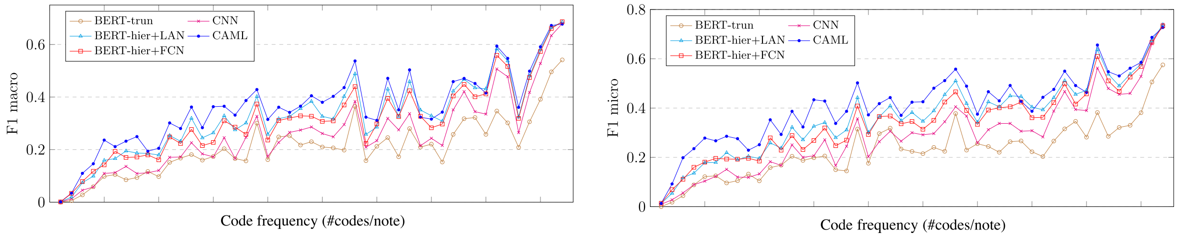
This section answers to the second research question (RQ2) by comparing the *BERT-trun* and *BERT-hier* fine-tuning architectures. Table 5 shows the results on the MIMIC-III top-50 code set using the two fine-tuning architectures when used either with a fully connected classifier or the label attention network. Our results are slightly better than the preliminary results reported by Ref. [10] with appropriate training tricks. Mixed domain pretraining models, such as PubMedBERT and ClinicalBERT, gain increases in evaluation scores. However, most *BERT-trun* variants' predictive performance suffers due to the lack of information when long notes are truncated into short ones. These results show that the proposed hierarchical fine-tuning architecture effectively utilizes long sequences and boosts performance. The label-wise attention mechanism can further improve the prediction in most cases.

4.4. *Reproducing CNN outperforms advanced methods on frequent codes (RQ3)*

Several recent advances for medical code assignment are CNN-based models. This section investigates whether improved training of CNN can

Table 6
Results of retrained CNNs on MIMIC-III with top-50 and full ICD codes. **Bold** text denotes the new state of the art obtained by an old model.

Model	MIMIC-III Top-50 Codes					MIMIC-III Full Codes					
	AUC-ROC		F1		P@5	AUC-ROC		F1		P@8	P@15
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro		
CNN (Kim, 2014)	87.6	90.7	57.6	62.5	62.0	80.6	96.9	4.2	41.9	58.1	44.3
CAML (Mullenbach et al., 2018)	87.5	90.9	53.2	61.4	60.9	89.5	98.6	8.8	53.9	70.9	44.5
MultiResCNN (Li and Yu, 2020)	89.9	92.8	60.6	67.0	64.1	91.0	98.6	8.5	55.2	73.4	58.4
HyperCore (Cao et al., 2020)	89.5	92.9	60.9	66.3	63.2	93.0	98.9	9.0	55.1	72.2	57.9
CNN (retrained)	**90.8**	**93.1**	**62.4**	**67.1**	64.0	85.0	97.4	5.9	36.5	49.0	39.4



(a) F1 macro

(b) F1 micro

Fig. 3. F1 scores of different models on the MIMIC-III full code dataset (8922 labels). Code frequency groups are sorted in ascending order from left to right.

improve prediction. Inspired by performance improvement owing to training tricks, we reproduce the CNN model to check if appropriate training strategies can lead to better scores.

4.4.1. Baselines

CNN [40] is built on pre-trained word vectors with 1D convolution and max-pooling for text classification. CAML [14] integrates CNNs and a label-wise attention mechanism to learn rich representations. It has a variant called DR-CAML that uses ICD code descriptions to regularize the loss function. MultiResCNN [9] combines residual learning [41] and multiple channels concatenation with different convolutional filters, achieving good performance in most settings. HyperCore [21] utilizes hyperbolic embedding and co-graph representation with code hierarchy. It gains slightly better performance than the MultiResCNN.

4.4.2. Reproducing CNNs

Table 6 shows the reproduced results compared with baselines. The hyper-parameters are as follows: the learning rate is 0.003, the filter size is 4, the number of filters is 500, the dropout probability is 0.2, and word2vec embeddings are static. Our retrained vanilla CNN gains a better performance compared with two recent advanced MultiResCNN [9] and HyperCore [21] on the MIMIC-III top-50 code dataset. Moreover, the P@15 score of the retrained CAML model leads to a significant performance increase in the MIMIC-III full code dataset. These results suggest that a simple model with appropriate training could achieve decent performance in NLP applications such as this clinical application.

How did the CNN- and BERT-based models perform on the full code set? We bin ICD codes into different groups according to the frequency observed in the training set of the MIMIC-III full code dataset to study the effect of code frequency on the model's predictive performance. We take ClinicalBERT as a clinical note encoding representative and compare three fine-tuning strategies with CNN and CAML model. Fig. 3 shows models' predictive behavior on frequent and less frequent codes taking F1 scores as the evaluation metrics. These two figures show that all compared models' predictive performance decreases when the ICD code groups have fewer training samples. When dealing with few-shot codes, model performance drops sharply. These results again suggest that medical code prediction algorithms should focus on less frequent codes and enhance the robustness of less frequent codes.

4.5. Discussion

Pretrained models in both general and specific domains have shown effectiveness in capturing contextual information. However, they encounter limitations in this study. In the dataset with top-50 frequent codes, fine-tuning with pretrained models can achieve a good performance; however, the PTM fine-tuning does not work well for high-dimensional structured prediction with a full label set that has more than 8000 labels. This study suggests focusing on less frequent codes.

Self attention-based models suffer from the complexity of $O(n^2d)$, where n is the sequence length, and d is the dimension of hidden representation, making it hard to encode extremely long documents. We investigated how to incorporate the pretrained BERT model and its variants with hierarchical fine-tuning architecture to tackle lengthy clinical document encoding. Nevertheless, CNN-based models [15,9,14] and RNN-based models [16] perform considerably well with the relatively small model scale and remain a meaningful direction. Recently, some improved transformer-based models such as Longformer [42], Linformer [43] and Big Bird [44] aim to solve the problem of encoding long document and mitigating the quadratic complexity. We leave these emerging models as future work.

5. Conclusion

This paper presented a comprehensive quantitive analysis of medical code assignment from clinical notes using various pretrained models with BERT. We compared the behavior of several different domain-specific BERT variants. To solve the problem of lengthy clinical note encoding, we developed two fine-tuning architectures: 1) fully connected network with simple truncation into short notes; 2) hierarchical fine-tuning architecture with long note segmentation and an additional Transformer on top. Moreover, we employed label attention to facilitate label-aware representation learning. Through intensive experiments, we found that the magic of BERT does not apply to the task of assigning ICD codes from clinical notes. In contrast, we found that a simple CNN trained from scratch can achieve superior predictive performance on frequent codes, achieving a new state of the art in the MIMIC-III top-50 dataset. This demonstrates how recent training strategies can improve old models. Our results furthermore suggest that medical code assignment algorithms should pay more attention to less frequent codes.

Author contributions

Conception and design of study: Shaoxiong Ji and Matti Hölttä; Analysis and/or interpretation of data: Shaoxiong Ji and Matti Hölttä; Drafting the manuscript: Shaoxiong Ji; Revising the manuscript: Shaoxiong Ji, Matti Hölttä and Pekka Marttinen.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the Academy of Finland (grant 336033) and EU H2020 (grant 101016775). We acknowledge the computational resources provided by the Aalto Science-IT project. The authors wish to acknowledge CSC - IT Center for Science, Finland, for computational resources.

References

[1] D. Zhang, J. Thadajarassiri, C. Sen, E. Rundensteiner, Time-aware transformer-based network for clinical notes series prediction, in: Machine Learning for Healthcare Conference, PMLR, 2020, pp. 566–588.

[2] D.C. Hsia, W.M. Krushat, A.B. Fagan, J.A. Tebbutt, R.P. Kusserow, Accuracy of diagnostic coding for medicare patients under the prospective-payment system, N. Engl. J. Med. 318 (6) (1988) 352–355.

[3] M.H. Stanfill, M. Williams, S.H. Fenton, R.A. Jenders, W.R. Hersh, A systematic literature review of automated clinical coding and classification systems, J. Am. Med. Inf. Assoc. 17 (6) (2010) 646–651.

[4] K. Crammer, M. Dredze, K. Ganchev, P. Talukdar, S. Carroll, Automatic code assignment to medical text, in: Biological, Translational, and Clinical Language Processing, 2007, pp. 129–136.

[5] R. Farkas, G. Szarvas, Automatic construction of rule-based ICD-9-CM coding systems, in: BMC Bioinformatics, vol. 9, Springer, 2008, pp. 1–9. Suppl 3.

[6] M. Farzandipour, A. Sheikhtaheri, F. Sadoughi, Effective factors on accuracy of principal diagnosis coding based on international classification of diseases, the 10th revision (ICD-10), Int. J. Inf. Manag. 30 (1) (2010) 78–84.

[7] M. Falis, M. Pajak, A. Lisowska, P. Schrempf, L. Deckers, S. Mikhael, S. Tsaftaris, A. O'Neil, Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text, in: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis, LOUHI 2019, 2019, pp. 168–177.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, in: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT, 2019.

[9] F. Li, H. Yu, ICD coding from clinical text using multi-filter residual convolutional neural network, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.

[10] S. Ji, S. Pan, P. Marttinen, Medical Code Assignment with Gated Convolution and Note-Code Interaction, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, 2021.

[11] H. Dong, V. Suárez-Paniagua, W. Whiteley, H. Wu, Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation, J. Biomed. Inf. Elsevier (2021).

[12] J. Medori, C. Fairon, Machine learning and features selection for semi-automatic ICD-9-CM encoding, in: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, Association for Computational Linguistics, 2010, pp. 84–89.

[13] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: models and evaluation metrics, J. Am. Med. Inf. Assoc. 21 (2) (2014) 231–237.

[14] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proceedings of NAACL-HLT, 2018, pp. 1101–1111.

[15] S. Ji, E. Cambria, P. Marttinen, Dilated convolutional attention network for medical code assignment from clinical text, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, 2020.

[16] H. Shi, P. Xie, Z. Hu, M. Zhang, E.P. Xing, Towards Automated ICD Coding Using Deep Learning, arXiv preprint arXiv:1711.04075, 2017.

[17] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, N. Elhadad, Multi-label classification of patient notes: case study on ICD code assignment, in: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[18] P. Xie, E. Xing, A neural architecture for automated icd coding, Long Papers, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, 2018, pp. 1066–1076.

[19] A. Prakash, S. Zhao, S.A. Hasan, V. Datla, K. Lee, A. Qadir, J. Liu, O. Farri, Condensed memory networks for clinical diagnostic inferencing, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[20] T. Bai, S. Vucetic, Improving medical code prediction from clinical text via incorporating online knowledge sources, in: The World Wide Web Conference, 2019, pp. 72–82.

[21] P. Cao, Y. Chen, K. Liu, J. Zhao, S. Liu, W. Chong, HyperCore: Hyperbolic and Co-graph representation for automatic ICD coding, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3105–3114.

[22] Q. Xipeng, S. TianXiang, X. Yige, S. Yunfan, D. Ning, H. Xuanjing, Pre-trained models for natural language processing: a survey, Sci. China Technol. Sci. 63 (2020) 1872–1897.

[23] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11 (2010) 625–660.

[24] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[25] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.

[26] A. Mulyar, E. Schumacher, M. Rouhizadeh, M. Dredze, Phenotyping of Clinical Notes with Improved Document Classification Models Using Contextualized Neural Language Models, arXiv preprint arXiv:1910.13664, 2019.

[27] K. Huang, J. Altosaar, R. Ranganath, ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, arXiv preprint arXiv:1904.05342, 2019.

[28] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, in: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, ACL, 2020.

[29] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 19–27.

[30] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 58–65.

[31] M. Basaldella, F. Liu, E. Shareghi, N. Collier, COMETA: A corpus for medical entity linking in the social media, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020.

[32] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific Language Model Pretraining for Biomedical Natural Language Processing, arXiv preprint arXiv:2007.15779, 2020.

[33] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment Pre-training for Biomedical Entity Representations, arXiv preprint arXiv:2010.11784, 2020.

[34] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, 2018, pp. 328–339. Long Papers.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[36] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019.

[37] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical Care database, Scientific Data 3 (2016) 160035.

[38] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.

[39] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[40] Y. Kim, Convolutional Neural Networks for Sentence Classification, arXiv preprint arXiv:1408.5882, 2014.

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[42] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The Long-Document Transformer, arXiv preprint arXiv:2004.05150, 2020.

[43] S. Wang, B. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-Attention with Linear Complexity, arXiv preprint arXiv:2006.04768, 2020.

[44] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big Bird: Transformers for Longer Sequences, arXiv preprint arXiv:2007.14062, 2020.