



Research article

Automated ICD coding for coronary heart diseases by a deep learning method



Shuai Zhao^{a,1}, Xiaolin Diao^{a,1}, Yun Xia^a, Yanni Huo^a, Meng Cui^b, Yuxin Wang^a, Jing Yuan^c, Wei Zhao^{d,*}

^a Department of Information Center, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100037, China

^b Medical Record Department, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100037, China

^c Department of Information Center, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100037, China

^d Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100037, China

ARTICLE INFO

Keywords:

ICD coding
Coronary heart diseases
Deep learning
BERT
Interpretability

ABSTRACT

Automated ICD coding via machine learning that focuses on some specific diseases has been a hot topic. As one of the leading causes of death, coronary heart diseases (CHD) have seldom been specifically studied by related research, probably due to lack of data concretely targeting at the diseases. Based on Fuwai-CHD and MIMIC-III-CHD, which are a private dataset from Fuwai Hospital and the CHD-related subset of a public dataset named MIMIC-III respectively, this study aimed at automated CHD coding by a deep learning method, which mainly consists of three modules. The first is a *BERT* variant module responsible for encoding clinical text. In the module, we fine-tuned *BERT* variants with masked language model on clinical text, and proposed a truncation method to tackle the problem that *BERT* variants generally cannot handle sequences containing more than 512 tokens. The second is a word2vec module for encoding code titles and the third is a label-attention module for integrating the embeddings of clinical text and code titles. In short, we named the method *BW_att*. We compared *BW_att* against some widely studied baselines, and found that *BW_att* performed best in most of the coding missions. Specifically, *BW_att* reached a *Macro-F1* of 96.2% and a *Macro-AUC* of 98.9% for the top-100 most frequent codes in Fuwai-CHD, which covered 89.2% of the total code occurrences. When predicting the top-50 most frequent codes in MIMIC-III-CHD, *BW_att* reached a *Macro-F1* of 40.5% and a *Macro-AUC* of 66.1%. Moreover, *BW_att* was capable of locating informative tokens from clinical text for predicting the target codes. In summary, *BW_att* can not only suggest CHD codes accurately, but also possess robust interpretability, hence has great potential in facilitating CHD coding in practice.

* Corresponding author.

E-mail address: zw@fuwai.com (W. Zhao).

¹ The authors contributed equally.

1. Introduction

ICD coding means assigning appropriate codes from the International Classification of Diseases system² to clinical text generated during patients' hospital stays. The coding task underlies many important applications, such as reimbursement and health service research [1,2]. At present, the task is mainly accomplished by clinical coders. As the ICD system is quite complicated and continuously updates, training a qualified coder is quite difficult. Besides, manual coding is costly in time and encounters problems such as inconsistency, making automated ICD coding through machine learning a popular topic.

Many studies targeted at auto-assigning all available ICD codes in a given dataset, and the typical ones are those predicting the over 8000 codes in MIMIC-III [3–5] by proposing deep learning models. These studies shed light on how to construct neural networks suitable for the ICD coding mission, however, all the proposed models are not accurate enough to be applied in practice.

Predicting all available ICD codes is quite challenging, as generally there is not enough training data for the whole ICD system. To promote the application of machine learning coding models, narrowing the target from all available codes to a subset of codes relating to some specific diseases is a common and effective method (such as [6,7]).

Coronary heart diseases (CHD) are among the leading causes of death. Over 100 million people worldwide lived with CHD, resulting in more than 8 million deaths [8]. Accurate CHD coding³ lies foundation for high quality data which can be used in many ways to prevent CHD, such as the development of diagnosis models [9,10]. However, to the best of our knowledge, automated CHD coding via machine learning has seldom been specifically studied, probably due to lack of datasets concretely targeting at the diseases, and this study aimed at achieving the goal based on two datasets. The first is Fuwai-CHD coming from Fuwai Hospital, which is a well known tertiary hospital for treating all kinds of complex cardiovascular diseases, and the second is MIMIC-III-CHD, which is the CHD-related subset of MIMIC-III.

Specifically, we accomplished automated CHD coding by proposing a deep learning method, which mainly comprises three modules. The first is a *BERT* [11] variant module for encoding clinical text. To adapt *BERT* variants to medical field, we fine-tuned them with masked language model (*MLM*) [11] on clinical text before using them in the coding missions. Moreover, we proposed a truncation method based on feature selection to tackle the problem that *BERT* variants generally can not handle input sequences containing more than 512 tokens. The second is a word2vec (*W2V*) module for encoding target code titles, as code titles can also provide useful information to assist code assignment [1]. The third is a label-attention module, which is inspired by existing deep learning coding models, and responsible for integrating the embeddings of clinical text and code titles. We name the method *BW_att* in short below. We assumed that promising CHD coding results could be achieved with *BW_att*, and served as a strong basis for assisting CHD coding in practice.

The contributions of the study lie in the following aspects. First, we achieved automated CHD coding via machine learning, which has received little attention so far. Second, we proposed *BW_att*, a new deep learning coding method that employs fine-tuned *BERT* variants as clinical text encoders, and overcomes the length limit of *BERT* variants by a new text truncation method. Third, we demonstrated the advantages of *BW_att* by comparing it with some widely studied baselines, and showed the coding decisions of *BW_att* were interpretable.

The rest of the paper is organized as follows. Section 2 describes our research data and methodology. Section 3 gives our experiment results. Section 4 analyzes the experiment results, the interpretability of *BW_att* and the limitations of this study. Section 5 concludes the whole paper.

2. Related work

2.1. Two directions of researches closely relate to this study

The first is developing deep learning models to predict all the diagnosis codes in MIMIC-III, which is a public dataset and contains admission records from the intensive care unit (ICU) of a US hospital. *CAML* [9] and *LAAT* [10] are two such models that have received much attention. *CAML* employs a label-attention layer stacked on a convolutional layer, and achieved a *Macro-F1* of 8.8%. Since proposed, *CAML* has commonly appeared as a baseline for subsequent models [12–20], such as *HyperCore* [12] that featured in mining code hierarchy and reached a *Macro-F1* of 9%, and *Fusion* [20] that extracted both global and local text features and resulted in a *Macro-F1* of 8.3%. *LAAT* contains a label-attention layer as a key component as well, but stacks it on a Bi-LSTM (bidirectional long short-term memory) [21] layer. Reaching a *Macro-F1* of 9.9%, *LAAT* has become one of the most promising models over recent years. Similar to *CAML* and *LAAT*, many other models, such as *G_coder* [14] and *MSATT_KG* [15], also adopt label-attention layers as critical structures, suggesting that the attention mechanism quite fits the ICD coding mission. Existing studies generally used convolutional or recurrent modules as clinical text encoders. Compared with these modules, large pretrained language models like *BERT* are more capable of analyzing syntax and semantics, hence have great potential in serving as better clinical text encoders. However, *BERT* variants can only handle a sequence of up to 512 tokens in general, which constrains their applications in analyzing quite long clinical text [22]. How to effectively employ *BERT* variants under the scenario of ICD coding without the length limit has not been extensively studied.

The second is predicting a subset of ICD codes that relate to some specific diseases or are of particular importance. As examples,

² The codes in the ICD system represent either diagnoses or procedures. We confine ICD coding to the assignment of diagnosis codes in this study.

³ To simplify expression, we use 'CHD coding' to represent ICD coding for CHD in this study.

Koopman et al. [6] employed support vector machine (SVM) to build a binary filter classifier, with the purpose of coding cancers in death certificates, and achieved a *F1*-score of 94.2%. Also using SVM, Karimi et al. [7] predicted 16 codes relating to radiology reports, reaching a *Micro-F1* of over 80%. Kaur et al. [23] compared various classifiers over predicting the codes associated with digestive and respiratory systems, and resulted in an optimal *F1*-score of 95%. Diao et al. [24] trained LightGBM models to identify the main diagnosis codes of clinical records and reached a *Macro-F1* of 88.3%. As far as we know, automated CHD coding via machine learning has seldom been specifically studied.

3. Materials and methods

3.1. Experiment design

Overall, our experiments were implemented as Fig. 1 displays. Before delving into the CHD coding mission, we first trained W2V embeddings and fine-tuned *BERT* variants based on the original datasets. Then, we selected CHD-related subsets from the original datasets. Finally, we addressed CHD coding via *BW_att*, and compared it against some widely studied baselines. Given the data of a specific CHD coding mission, 70% was used for training, 10% was for validation and the rest was for test.

3.2. Data

Fuwai dataset. Approved by the Ethics Committee at Fuwai Hospital, we gained a private Chinese dataset which lasts from January 2019 to December 2020, in which no identifiable personal data is included. The dataset involves over 75,000 admission records. Each record contains a textual diagnosis summary, which is mainly about diagnoses and treatment process, and ground truth ICD-10 disease codes and ICD-9 procedure codes.

MIMIC-III. MIMIC-III is a public English dataset from the Beth Israel Deaconess Medical Center in the US, and contains over 58,000 admission records of patients in the ICU between 2001 and 2012. Each record contains a textual discharge summary, which includes information about diagnoses and medication, and ground truth ICD-9 disease and procedure codes.

Training W2V and fine-tuning BERT variants. As preprocessing on the clinical text in both datasets, we turned all English tokens from uppercase into lowercase, and removed all numbers and symbols like related studies did [9,10,15].

We extracted sentences from all the records in Fuwai dataset and MIMIC-III respectively. For Fuwai dataset, we used *Jieba*⁴ package to cut the sentences into word sequences. To identify medical terms more accurately, we added a medical vocabulary provided by *Sogou*⁵ into *Jieba*, which includes 90,047 terms pertaining to diagnoses and so on. For MIMIC-III, we added all the long code titles into the sentence set. Subsequently, we trained W2V embeddings based on the sentences, and the parameters of the W2V models as well as their descriptions are given in Table 1.

In terms of the *BERT* variants in *BW_att*, for Fuwai dataset, there are both character-based and word-based Chinese pre-trained models available. However, word-based pre-trained models, like *WoBert* [25], generally cover a limited number of words, and would run into serious out-of-vocabulary problems when used in analyzing medical text. Therefore, we adopted a character-based pre-trained model called *RoBERTa-Mini* [11,26–28], which has four transformer encoder layers and outputs character embeddings with 256 dimensions in each layer. For MIMIC-III, we employed an English pre-trained model⁶ called *BERT-mini* [29,30], whose network structure is the same as that of *RoBERTa-Mini*.

Both *RoBERTa-Mini* and *BERT-mini* were pre-trained on a large amount of multi-domain text. With the purpose of enhancing their capability in mining clinically useful information, we fine-tuned them with *MLM* for 20 epochs, based on the sentences from Fuwai dataset and MIMIC-III respectively, and used them for code assignment afterwards.

Data selection for CHD. Following the professional advice from the physicians in Fuwai Hospital, we selected the records relating to CHD by setting some rules on the procedures undertaken by the patients, with the purpose of ensuring that the main treatment target of each selected record is CHD-related diseases. To save some space, we give the details of the rules in Supplementary File 1. Some descriptive statistics of Fuwai-CHD and MIMIC-III-CHD, which separately stand for the CHD-related subsets of Fuwai dataset and MIMIC-III, are listed in Table 2.

To figure out how the diagnosis codes distribute among Fuwai-CHD and MIMIC-III-CHD, we sorted the codes according to frequency in descending order, and showed how the frequencies change along with the rankings in Fig. 2(a)–(b). The figures display two apparent long-tail distributions, with the top 5% most frequent codes accounting for 89% and 75% of the total code occurrences in Fuwai-CHD and MIMIC-III-CHD, respectively. This indicates that even if machine learning models merely covered the frequent codes, the models could still potentially alleviate a large portion of the burden of clinical coders. Hence, we mainly targeted at the frequent codes in this study, as the codes with low frequency would lead machine learning models to the risk of overfitting.

3.3. BW_att

Fig. 3 shows the diagram of *BW_att*, which mainly comprises three modules, a *BERT* variant module, a W2V module [31], and a

⁴ <https://pypi.org/project/jieba/>.

⁵ Available at: <https://pinyin.sogou.com/dict/cate/index/132>.

⁶ English pre-trained models are generally word-based.

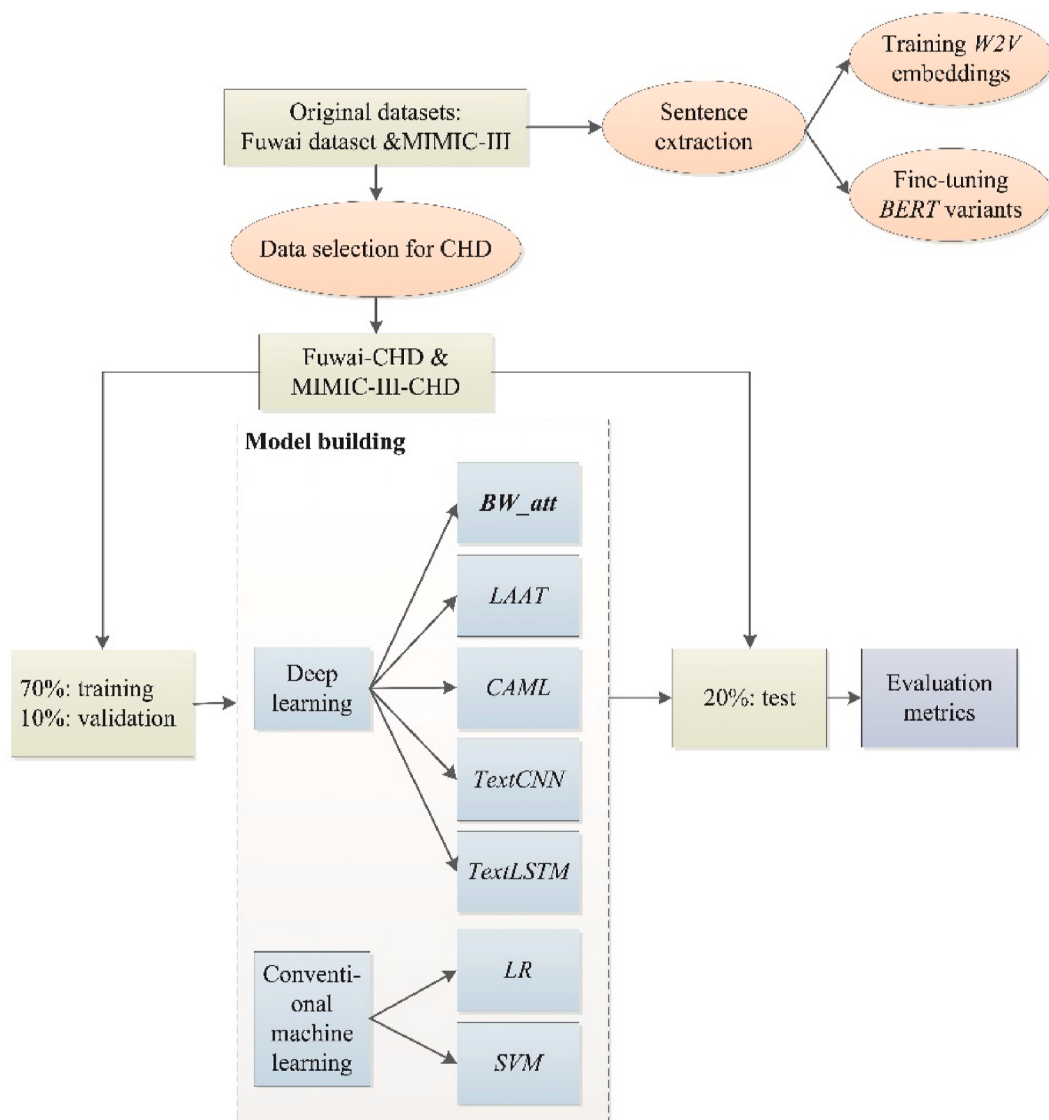


Fig. 1. The main design of our experiments.

Table 1

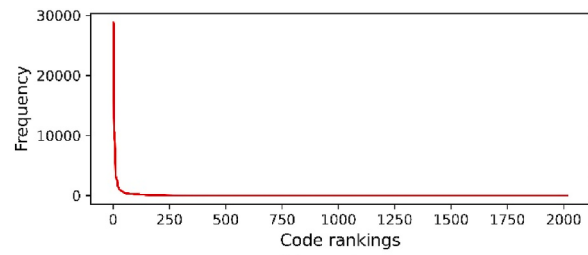
The parameters of the *W2V* models. 16-char and 16-word represent char-level and word-level embeddings with 16 dimensions, respectively. For Fuwai-CHD, 128-word embeddings were used to initialize the embedding layers of the deep learning baselines, and the rest were used for optimizing the hyper-parameters of *BW_att*. For MIMIC-III-CHD, 64-word embeddings were used to initialize the embedding layers of the deep learning baselines, and all the embeddings were used to optimizing the hyper-parameters of *BW_att*.

Parameters	Descriptions	For Fuwai-CHD	For MIMIC-III-CHD
sg	Whether Skip-gram is used or not.	1	1
dimension-type	The type and dimension of resulting embeddings.	16-char, 32-char, 64-char, 16-word, 32-word, 64-word, 128-word	32-word, 64-word, 128-word
window	The length of a text window.	5	5
iter	Training for how many iterations.	5	5
min_count	Discarding tokens appearing less than how many times.	3	3

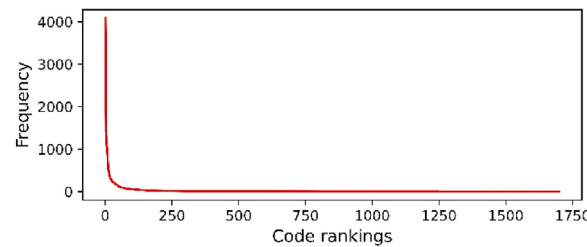
Table 2

Some descriptive statistics of Fuwai-CHD and MIMIC-III-CHD. A token means a character in Fuwai-CHD and a word in MIMIC-III-CHD. Token length above 512 indicates the number of records with clinical text longer than 512 tokens.

	Fuwai-CHD	MIMIC-III-CHD
Number of records	31,052	4361
Token size	6,418,316	3,316,405
Token vocabulary size	1890	35,365
Average token length	206.7	760.5
Token length above 512	418	3314
Code size	221,532	38,595
Code vocabulary size	2016	1700
Average code length	7.1	8.9



(a) Fuwai-CHD



(b) MIMIC-III-CHD

Fig. 2. The distributions of the code frequencies in the CHD-related datasets. Panels (a) and (b) are for Fuwai-CHD and MIMIC-III-CHD respectively.

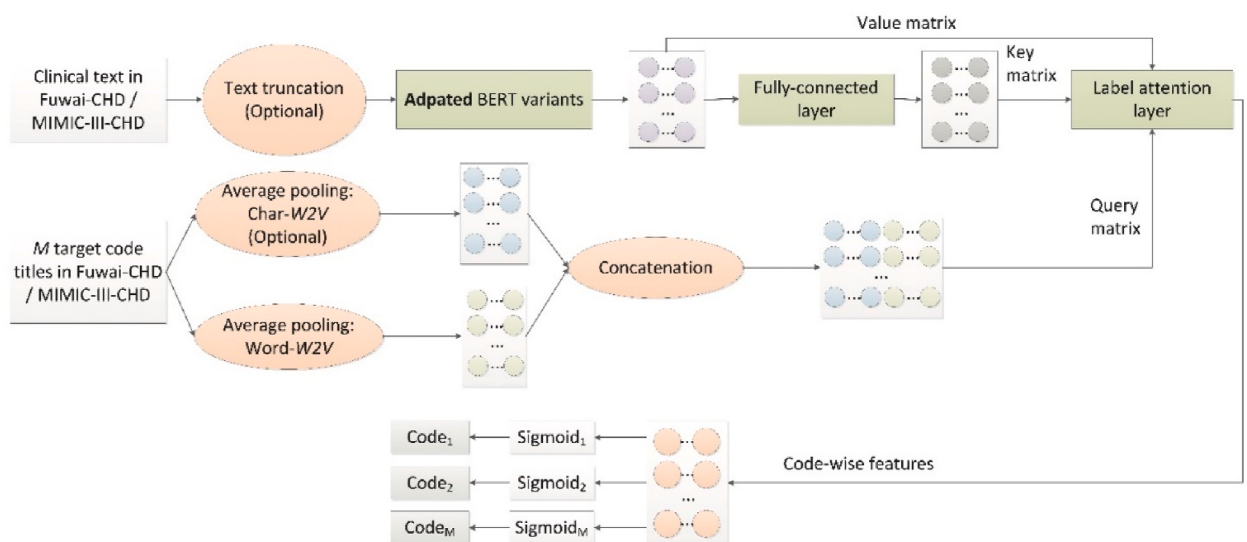


Fig. 3. The diagram of BW_{att} .

label-attention module. Details of the modules are described next.

BERT variant module. Given an input sequence with n tokens $s = \{c_1, c_2, \dots, c_n\}$, we denoted the output embedding matrix⁷ of the BERT variants as $V_s \in R^{n \times rdim}$, which was computed with Equation (1):

$$V_s = \text{BER_var}(s) \quad (1)$$

With $rdim = 256$.

Text truncation. The mainstream of BERT variants can handle a sequence of up to 512 tokens, and sequences longer need to be truncated. One commonly used truncation strategy is the truncation at back or front, which might drop useful information in ICD coding missions. To resolve the problem, we proposed a new truncation method, whose procedures are shown in Fig. 4.

Using the method, token features are extracted with bag-of-words and term frequency-inverse document frequency (tf-idf) from training data in the first place. Then, feature selection with filter methods are implemented to select important tokens. Finally, training, validation and test data was truncated by deserting unimportant tokens.

Regarding the filter method, we utilized Chi-square test to calculate the contribution of each candidate token with respect to assigning each target code. Among the multiple contributions associated with multiple target codes, the largest is used as the final metric for the contribution. After ranking all candidate tokens in descending order according to the contribution, the top- n were selected as important tokens.

Note that a smaller n means more severe truncation. There does not exist a single n that is universally appropriate, as factors like language can significantly affect what value n should take. Moreover, there is no need to adopt the truncation method under the condition that clinical text in most records contains less than 512 tokens.

As Table 2 shows, over 98.5% of the diagnosis summaries in Fuwai-CHD contain less than 512 characters, so we just conducted the truncation at back⁸ on the dataset when necessary. Regarding MIMIC-III-CHD, around 76% of the discharge summaries are longer than 512 words, and thus we adopted our truncation method on its records throughout our experiments.

W2V module. In the Chinese context, existing research concluded that information from both words and characters is important and tends to be complementary in analyzing clinical text [32]. Accordingly, We encoded the code titles in Fuwai-CHD using both the word and character embeddings. In specific, given a code c_i , we obtained the embeddings q_i for its title with Equation (2):

$$q_i = \text{concat}(\text{ave_char_w2v}(c_i), \text{ave_word_w2v}(c_i)) \quad (2)$$

Where $\text{ave_char_w2v}(c_i)$ and $\text{ave_word_w2v}(c_i)$ stand for the average-pooling results over the embeddings of all the words and characters⁹ in the title respectively, and $\text{concat}(\cdot)$ means concatenation operation.

To encode the code titles in MIMIC-III-CHD, we did not use character embeddings, as in the English context, it is unlikely that characters like 'a' or 'b' directly involve useful information for code prediction. We merely used word embeddings and obtained q_i with Equation (3):

$$q_i = \text{ave_word_w2v}(c_i) \quad (3)$$

Label-attention module and output layer. The label-attention module is responsible for integrating the embeddings of clinical text and target code titles, and generating code-wise predictive features.

In the label-attention mechanism, we computed the key matrix using Equation (4):

$$K_s = V_s \cdot W_k \quad (4)$$

Where $W_k \in R^{rdim \times adim}$ is a trainable matrix. V_s is the embedding matrix for an input sequence s outputted by the BERT variant module, and also used as the value matrix.

For code c_i , we used its embeddings q_i as the query vector, and obtained its predictive features $f_i \in R^{1 \times rdim}$ with Equation (5):

$$\begin{cases} f_i = att_i \bullet V_s \\ att_i = \text{softmax}(q_i \bullet K_s) \end{cases} \quad (5)$$

where $att_i \in R^{1 \times n}$ contains attention weights for all input tokens in relation to the prediction of c_i .

Ultimately, the probability p_{si} that s should be assigned c_i was computed via a fully-connected layer as Equation (6) shows:

$$p_{si} = \text{sigmoid}(f_i \bullet W_i + b_i) \quad (6)$$

where $W_i \in R^{rdim \times 1}$ and b_i are trainable weights and bias respectively. Binary cross-entropy in Equation (7) was used as the loss function:

⁷ V_s does not contain the embeddings for [CLS] and [SEP] in this study.

⁸ As stated in 'Ablation study', we experimented with both the truncation at back and the truncation at, and found that the former generally led to better results.

⁹ The W2V embeddings are trained based on the sentences from Fuwai dataset and MIMIC- III.

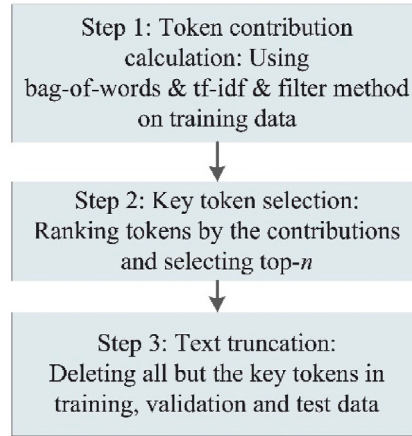


Fig. 4. The procedures of our text truncation method.

$$L = -\frac{1}{|S|} \sum_{s \in S} \sum_{i=1}^N [y_{si} \log p_{si} + (1 - y_{si}) \log(1 - p_{si})] \quad (7)$$

where S is a set of training data, N is the number of target CHD codes, and y_{si} equals 0 or 1 indicating whether c_i truly accompanies s or not.

3.4. Baselines

CAML & LAAT. As stated in Related work, both *CAML* and *LAAT* are typical representatives of deep learning coding models over the recent years, the main structures of which are shown in Fig. 5(a) and (b) respectively.

TextCNN & TextLSTM. *TextCNN* [33] (text convolutional neural network) and *TextLSTM* were commonly used as deep learning baselines in related studies (such as [9,10]). In terms of *TextLSTM*, we adopted an *LSTM* layer on text embeddings, and the output corresponding to the last token of an input sequence was fed into a fully-connected layer, which led to coding results.

LR & SVM. Among conventional machine learning models, *LR* is popular for its simplicity and high efficiency, while *SVM* is considered one of the best conventional text classifiers [34,35]. Both models are popular baselines in existing studies [6,23,36,37].

3.5. Experiment details

We implemented ablation study to validate the effectiveness of our text truncation method.

Regarding each coding mission, we tuned some hyper-parameters of the models, and the parameters leading to the lowest validation loss were considered optimal. The parameter searching spaces and prefixed hyper-parameters of all the models are listed in Supplementary File 2.

For the deep learning models, taking into account the randomness caused by factors like parameter initialization, we implemented five rounds of model building, with different rounds adopting different random seeds for data split. Mean metrics over the results of all the rounds of experiments were given finally.

One-sided Wilcoxon test was used to evaluate the robustness of the advantages of best-performing models against some selected baselines.

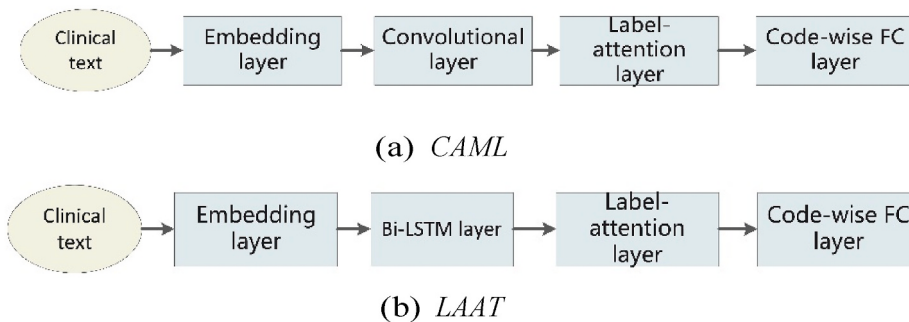


Fig. 5. The main structures of typical existing coding models. Panels (a) and (b) represent *CAML* and *LAAT* respectively.

3.6. Metrics

3.6.1. Performance metrics

In line with related studies [6,9,23,38,39], we mainly adopted *F1*-score and *AUC* to evaluate coding performance. *Micro-F1*, *Macro-F1*, *Micro-AUC* and *Macro-AUC* were used as specific metrics. Note that micro metrics focus more on codes with more data while macro metrics allocate same weights for all codes. We paid more attention to macro metrics in this study, as each code is equally important in coding practice.

F1-score and *AUC* in regards of a single code are introduced below.

F1-score. Assume there are some test records within which t_1 are in class 1 meaning a code is assigned. After being analyzed by a trained model, p_1 records are labeled as 1, among which tp_1 are accurately labeled. Then, the *F1*-score for the coding results is calculated with Equation (8):

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

where $\text{Precision} = tp_1/p_1$ and $\text{Recall} = tp_1/t_1$.

AUC. *AUC* stands for area under the curve, with the curve indicating receiver operating characteristic (ROC) curve. Many classification models output the probabilities that records belong to class 1, instead of directly 1 or 0. Hence, a calibration threshold is needed to transform the probabilities to classes. In terms of *ROC*, a number of such thresholds are first chosen, and true positive rate (*TPR*) and false positive rate (*FPR*) corresponding to each of the thresholds are then computed. Finally, the curve is generated by plotting the *TPRs* against the *FPRs*.

3.6.2. The metric for the explainability of *BW_att*

Interpretability is quite important for clinical applications of machine learning coding models. With *BW_att*, interpretability can be demonstrated by the attention weights from the label-attention module, with higher weights representing more importance of corresponding tokens for code assignment. By unveiling the distribution of the weights, we can easily observe what are key tokens and whether the key tokens are useful after referring to target code titles.

To quantify the explainability of *BW_att*, we defined *Pro_K* in Fig. 6, the idea behind which was to calculate the proportion of the *top-K* key tokens identified by *BW_att* that are truly useful in predicting target codes. A larger *Pro_K* indicates better explainability of *BW_att*.

4. Results

We selected the most frequent top-*N* codes¹⁰ in Fuwai-CHD and MIMIC-III-CHD to predict, and only the records involving at least one of the selected codes were used for model building.

4.1. For multiple top-*N* codes

We let *N* take 60, 80, 100, 120, 140 for Fuwai-CHD, and 20, 30, 40, 50, 60 for MIMIC-III-CHD.¹¹ The statistics about the subset of data with each *N* are given in Table 3.

When implementing our text truncation method on MIMIC-III-CHD, we empirically fixed the number of the key tokens as 2000 throughout our experiments. Taking the subset for the top-50 codes as an example, we give some statistics about the length of the discharge summaries before and after the truncation in Table 4.

The optimal hyper-parameters and detailed coding performance of the models are given in Supplementary File 3. Fig. 7 and Fig. 8¹² show the *AUC* and *F1*-scores of the models on Fuwai-CHD and MIMIC-III-CHD, respectively.

We learnt from Fig. 7 that *BW_att* and *LAAT* performed better than the rest of the models in terms of *Macro-F1*. When *N* equalled or exceeded 100, *BW_att* performed best. Regarding the top-140 codes, *BW_att* achieved a *Macro-F1* of 95.4%, a *Macro-Precision* of 93.3% and a *Macro-Recall* of 97.9%, and its advantage over *LAAT* regarding *Macro-F1* was statistically significant at the significance level of 10%¹³ (*P*-value = 0.094). When *N* fell below 100, *LAAT* significantly outperformed *BW_att* in terms of *Macro-F1* (with *P*-values of 0.03), and led to a *Macro-F1* of 98.1%, a *Macro-Precision* of 97.3% and a *Macro-Recall* of 98.9% for the top-60 codes. Aside from *LAAT* and *BW_att*, *CAML* also showed competitive performance.

Regarding MIMIC-III-CHD, except for *Micro-F1* when *N* equalled 20 or 40, *BW_att* turned out to be the best-performing model across all the values of *N* over all the metrics. The advantages of *BW_att* regarding *Macro-F1* over *CAML*, which was the second-best model, were consistently significant (with *P*-values of 0.03), and tended to enlarged as *N* grew. For the top-60 codes, *BW_att* resulted in a *Macro-F1* of 37.5%, a *Macro-Precision* of 42.2% and a *Macro-Recall* of 35.6%. Considering the other models, *SVM* was the most promising one.

¹⁰ We simplified 'the most frequent top-*N* codes' as 'the top-*N* codes' below.

¹¹ We chose the values of *N* according to the standard that the least frequent qualified code involves at least (or very close to) 100 records.

¹² All the panels in Fig. 7 share the same legend, and this is the same for Figs. 8 and 9.

¹³ The *P*-values in terms of all the metrics were provided in Supplementary File 5.

Algorithm: Calculate Pro_K to evaluate the interpretability of BW_att on Fuwai-CHD.	
Input: A test set with n clinical records	
Output: Pro_K	
For test record $i=1, 2, \dots, n$:	
1. Obtain top K key tokens from diagnosis summary with BW_att ;	
2. Confirm m_i , the size of the intersection of the key tokens and target code title tokens;	
3. $Pro_K_i = \frac{m_i}{K}$;	
$Pro_K = \frac{1}{n} \sum_{i=1}^n Pro_K_i$	
Return Pro_K	

Fig. 6. The algorithm for calculating Pro_K .

Table 3

Some descriptive statistics of the datasets in the coding missions. Proportion taken by the codes means the percentage of the total code frequency covered by the top- N codes.

	top- N	Number of records	Code occurrences	Proportion taken by the codes
MIMIC-III-CHD	top-20	4349	19,713	51.1%
	top-30	4352	22,246	57.6%
	top-40	4354	24,188	62.7%
	top-50	4356	25,678	66.5%
	top-60	4359	26,756	69.3%
Fuwai-CHD	top-60	31,011	186,687	84.3%
	top-80	31,015	192,982	87.1%
	top-100	31,039	197,572	89.2%
	top-120	31,040	201,206	90.8%
	top-140	31,043	204,039	92.1%

Table 4

Some statistics about the token length of the records in MIMIC-III-CHD. As for the truncated discharge summaries still longer than 512 words, we additionally implemented the truncation at back.

	Training		Validation		Test	
	Original	Truncated	Original	Truncated	Original	Truncated
Number of records	3049	3049	436	436	871	871
Token size	2,313,831	822,084	327,918	116,309	670,726	238,412
Token vocabulary size	29,306	2000	11,514	949	16,299	1112
Average token length	758.9	269.6	752.1	266.8	770.1	273.7
Token length above 512	2305	112	340	14	665	31

In summary, BW_att performed better than the baselines in most of the coding missions.

Besides the prediction results, we also recorded the running time of BW_att . All of our experiments were accomplished using a NVIDIA Tesla V100 GPU, and Table 5 lists the training and test time of BW_att across the coding missions. The results showed that all the model training was finished within 40 and 8 min for the top- N codes in Fuwai-CHD and MIMIC-III-CHD respectively. There are some training time outliers, such as the one for the top-40 codes in MIMIC-III-CHD, which is longer than the training time for the top-60 codes. This is probably due to the random issues induced by factors like parameter initialization and early stop training strategy. As for the test phase, one single record in either Fuwai-CHD or MIMIC-III-CHD could be handled in less than 0.01 s.

4.2. Ablation study

Replacing our text truncation method with the truncation at back and front respectively, we rebuilt BW_att on MIMIC-III-CHD to predict the multiple top- N codes. Under each N , we used the same hyper-parameters as those in Section 4.1.

Fig. 9 depicts the performances of BW_att with the three truncation methods, which showed that BW_att with our method performed

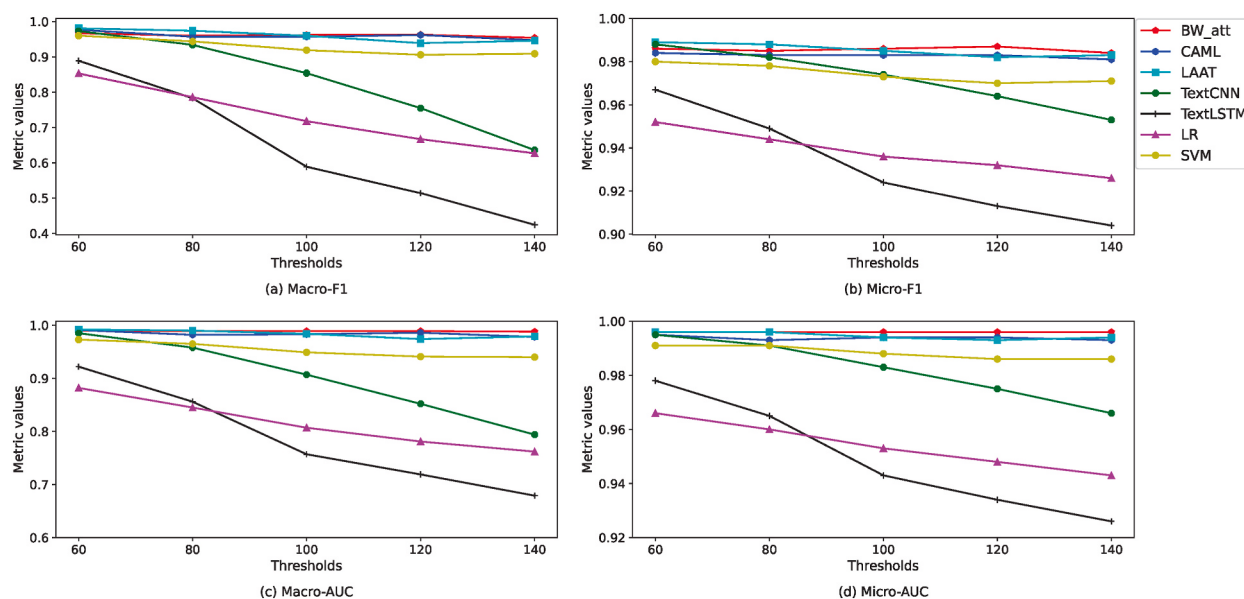


Fig. 7. The performance of the models on predicting the multiple top- N codes in Fuwai-CHD. Panels (a), (b), (c) and (d) display *Macro-F1*, *Micro-F1*, *Macro-AUC* and *Micro-AUC* of the models respectively.

Table 5

The running time of *BW_att* across all the coding missions. The training and test time is the average over that from all the rounds of experiments, and the test time is for all the test records in a coding mission.

MIMIC-III-CHD			MIMIC-III-CHD		
Thresholds	Training time (mins)	Test time (secs)	Thresholds	Training time (mins)	Test time (secs)
60	34.6	7	20	4.7	1
80	31.1	7	30	5.3	1
100	33.4	6.6	40	7.1	1
120	35.4	6.2	50	6.3	1
140	39.5	7	60	6.7	1

significantly better than those with the other two methods (with P -values of 0.03). Compared with the truncation at front, the truncation at back led to better results. The conclusions confirm that our text truncation method is effective in retaining useful information for code assignment.

5. Discussion

5.1. Result analysis

The experiments showed that all the models performed more promisingly on Fuwai-CHD than on MIMIC-III-CHD, and this could be attributable to two reasons. First, the records in MIMIC-III-CHD are much less than those in Fuwai-CHD, so the model training based on Fuwai-CHD was more sufficient. Second, most codes in Fuwai-CHD represent specified diseases, such as I25.105 (coronary atherosclerotic heart disease). Tokens quite predictive of such codes (like ‘atherosclerotic’ for I25.105) could be electively located by machine learning

models, and as a result, good coding performance could be reached. In contrast, many codes in MIMIC-III-CHD stand for unspecified diseases, such as 4019 (unspecified essential hypertension) and 2724 (other and unspecified hyperlipidemia). Among the top-30 codes in MIMIC-III-CHD, 19 are of this kind (Please see the code list in Supplementary File 4.). The uncertainty behind such codes hinders the models from capturing solid patterns for forecasting the codes.

Focusing on Fuwai-CHD, *BW_att* turned out to be the best in the coding missions covering more relatively infrequent codes (top-100 or more), probably because *BW_att* was better at mining more informative features from the diagnosis summaries and code titles for those less frequent target codes. From the perspective of Wilcoxon test, the advantages of *BW_att* over *LAAT* appeared not quite significant. However, the performance metrics of all the models were the average over those from the experiments with different random seeds, which effectively lowered the randomness behind the coding results, and made the advantages of *BW_att* reliable.

As for MIMIC-III-CHD, *BW_att* consistently served as the best model. What’s more, the advantages of *BW_att* over the other models

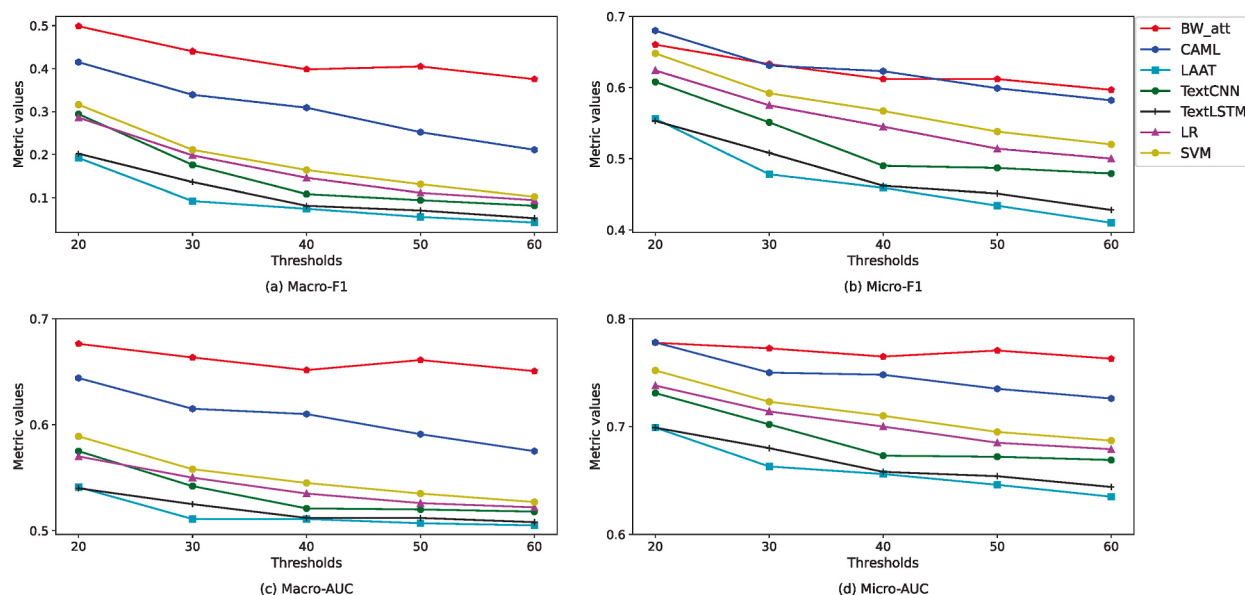


Fig. 8. The performance of the models on predicting the multiple top-N codes in MIMIC-III-CHD. Panels (a), (b), (c) and (d) display *Macro-F1*, *Micro-F1*, *Macro-AUC* and *Micro-AUC* of the models respectively.

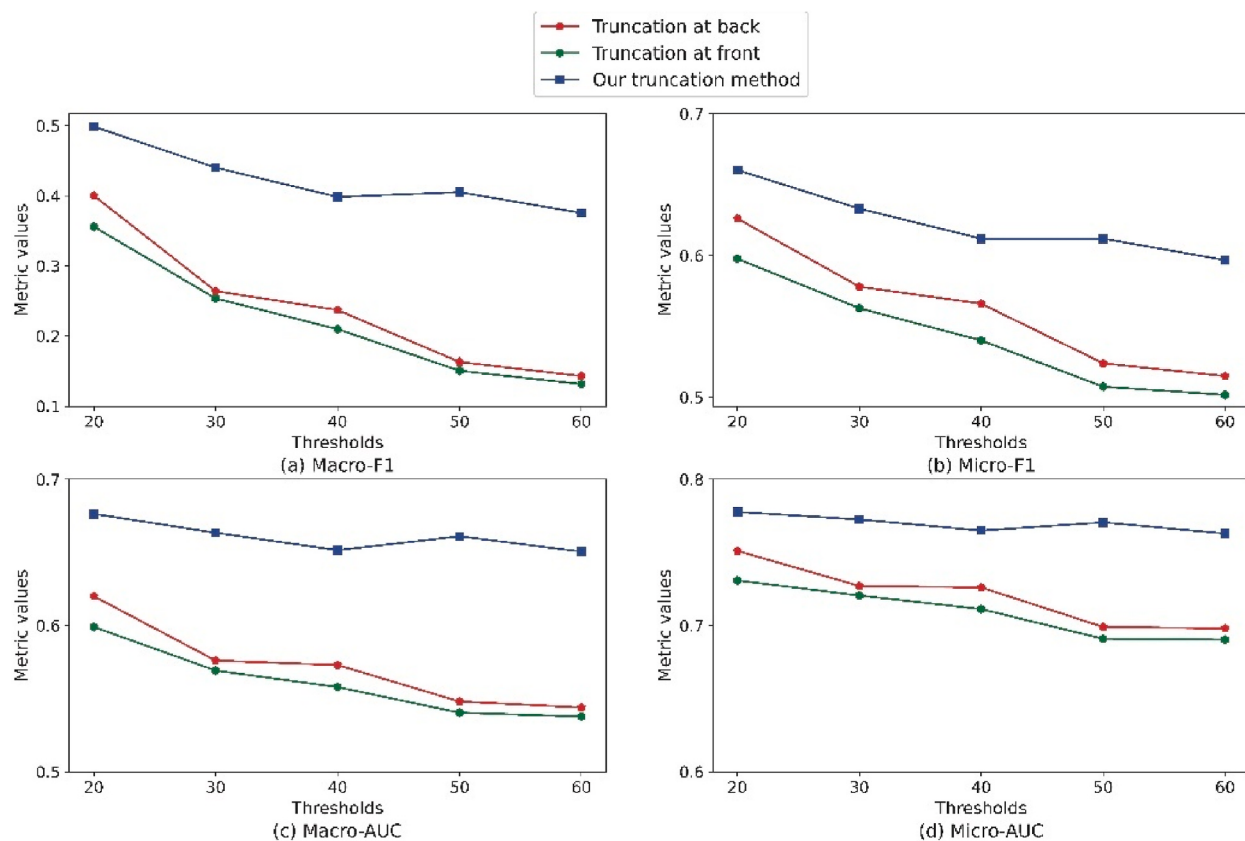


Fig. 9. The performances of *BW_att* with our text truncation method and the truncation at back and front on MIMIC-III-CHD. Panels (a), (b), (c) and (d) display *Macro-F1*, *Micro-F1*, *Macro-AUC* and *Micro-AUC* of the models respectively.

were more obvious than those based on Fuwai-CHD, and the main reason could be as follows. After being pretrained and fine-tuned, *BERT-mini* (embedded in *BW_att*) became capable of dealing with ambiguity in clinical text effectively, and this capability is of great help in extracting useful features to predict the codes representing unspecified diseases. Note that in all the coding missions, *BW_att* achieved significantly better performance than that reached by *LAAT* on predicting all the codes in MIMIC-III (*Macro-F1* = 9.9%), and this shows that focusing on the codes relating to CHD indeed improved the chance of applying the coding models in practice.

5.2. Interpretability

Regarding Fuwai-CHD, focusing on the top-100 codes as an example, we list *Pro_K* with different *K* calculated based on a randomly picked round of experiment in Table 6, which reflects that most of the key characters located by *BW_att* are directly indicative of the target codes. In other words, *BW_att* can assist CHD coding by suggesting the codes accurately (with a *Macro-F1* of 96.2%) and highlighting key text snippets at the same time. Note that the occurrences of the top-100 codes accounted for 89.2% of the total code occurrences, and auto-assigning these codes can effectively ease the burden of clinical coders and improve coding efficiency.

In terms of MIMIC-III-CHD, *Pro_K* does not fit for evaluating the explainability of *BW_att*. As mentioned in Section 5.1, many codes in MIMIC-III-CHD stand for unspecified diseases. Accordingly, certain key words identified by *BW_att* might not emerge in the target code titles, even though they are helpful in predicting the codes. Below, we show the interpretability of *BW_att* by giving a randomly selected record from the coding mission for the top-50 codes.

Table 7 lists the discharge summary and ICD codes of the record. The key words in the summary identified by *BW_att* are shown in bold. Among the keywords, only ‘mitral’ appears in the target code titles. However, other keywords also closely relate to the target codes. For instance, ‘counadin’ is a kind of medicine commonly used to treat diseases relating to valve (4240). ‘Chf’ stands for congestive heart failure, which is directly indicating code 4280. ‘Osa’ stands for obstructive sleep apnea, and is one of the hazard factors leading to coronary heart diseases (41,401).

5.3. Limitations

This study faces some limitations. First, we only experimented on a Chinese dataset and an English dataset. As recorded by existing studies [40,41], the portability of automated ICD coding models could not be guaranteed. In the future, we will validate the robustness of *BW_att* by experimenting on more datasets from different hospitals.

Second, in order to avoid overfitting, *BW_att* did not cover the rare codes relating to CHD. As the rare codes are also important in coding practice, we will make efforts to auto-assign them in the future. One research direction will be updating *BW_att* by adding a graph neural network module, with the purpose of taking correlations between frequent codes and rare codes into account. According to some related studies [15,42,43], such correlations could help forecast rare codes. Another direction will be setting some rules to identify the rare codes. Using *BW_att* and the rules together, more comprehensive automated CHD coding could be achieved.

6. Conclusion

CHD poses life threats to a great many people worldwide. Automated CHD coding via machine learning is of great help in laying data foundation for studying and preventing CHD, but has seldom been specifically studied. This paper aimed at achieving automated CHD coding by a deep learning method named *BW_att*. Experiments demonstrated that *BW_att* outperformed some widely studied baselines in most of the coding missions, reaching a *Macro-F1* of 96.2% for the top-100 codes in Fuwai-CHD, and a *Macro-F1* of 40.5% for the top-50 codes in MIMIC-III-CHD. Moreover, *BW_att* was capable of locating informative tokens from clinical text for predicting the target codes. Overall, *BW_att* has great potential in facilitating CHD coding in practice.

Author contribution statement

Shuai Zhao: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Xiaolin Diao: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Yun Xia: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Yanni Huo; Yuxin Wang: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Jing Yuan; Wei Zhao; Meng Cui: Conceived and designed the experiments; Wrote the paper.

Table 6
Pro_K for the interpretability of *BW_att* on Fuwai-CHD.

<i>K</i>	6	7	8	9	10
<i>Pro_K</i>	84.7%	84.0%	83.1%	82.2%	81.2%

Table 7A record for interpreting *BW_att* on MIMIC-III-CHD.

The discharge summary	ICD codes
.....bilaterally neuro pertinent results pm glucose urea creat sodium potassium chloride total anion echocardiogram impression intracardiac thrombus moderate severe mitral regurgitationphos mg hospital chf exacerbation displayed pulmonary edemadisplayed rapid afib admission displaying occasional episodes bradycardiathought osa pulm htn underwent hearttherapeutic coumadin admission prep left heartmg tablet delayed release ec sig	41,401 (coronary atherosclerosis of native coronary artery) 42,731 (atrial fibrillation) 4280 (congestive heart failure unspecified) 4240 (mitral valve disorders) 4168 (other chronic pulmonary heart diseases)

Funding statement

This work was supported by CAMS Innovation Fund for Medical Sciences [2018-I2M-AI-006].

Data availability statement

The data that has been used is confidential.

Declaration of interest's statement

The authors declare no competing interests.

Acknowledgments

We show our gratitude to the physicians in Fuwai Hospital who helped us select CHD-related records from the raw datasets.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.heliyon.2023.e14037>.

References

- [1] Y. Wu, M. Zeng, Z. Fei, Y. Yu, F.-X. Wu, M. Li, KAICD: a knowledge attention-based deep learning framework for automatic icd coding, *Neurocomputing* 469 (2022) 376–383.
- [2] L. Cao, D. Gu, Y. Ni, G. Xie, Automatic ICD code assignment based on ICD's hierarchy structure for Chinese electronic medical records, *AMIA Summits on Translational Sci.* 2019 (2019) 417–424.
- [3] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, *Circulation* 101 (23) (2000) e215–e220.
- [4] A. Johnson, T. Pollard, R. Mark, MIMIC-III clinical database (version 1.4), *PhysioNet* doi:<https://doi.org/10.13026/C2XW26>.
- [5] 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, *Sci. Data* 3 (1) (2016) 1–9.
- [6] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, N. Grayson, Automatic ICD-10 classification of cancers from free-text death certificates, *Int. J. Med. Inf.* 84 (11) (2015) 956–965.
- [7] S. Karimi, X. Dai, H. Hassanzadeh, A. Nguyen, Automatic Diagnosis Coding of Radiology Reports: A Comparison of Deep Learning and Conventional Classification Methods, *BioNLP*, 2017, pp. 328–332.
- [8] H. Dai, A.A. Much, E. Maor, E. Asher, A. Younis, Y. Xu, Y. Lu, X. Liu, J. Shu, N.L. Bragazzi, Global, regional, and national burden of ischaemic heart disease and its attributable risk factors, 1990–2017: results from the global burden of disease study 2017, *European Heart J. Qual. Care and Clinical Out.* 8 (1) (2022) 50–60.
- [9] J. Mullenbach, S. Wiegrefe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 1101–1111.
- [10] T. Vu, D.Q. Nguyen, A. Nguyen, A label attention model for ICD coding from clinical text, in: *International Joint Conference on Artificial Intelligence*, 2020, pp. 3335–3341.
- [11] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018 arXiv e-prints arXiv: 1810.04805.
- [12] P. Cao, Y. Chen, K. Liu, J. Zhao, S. Liu, W. Chong, Hypercore: hyperbolic and co-graph representation for automatic ICD coding, in: *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3105–3114.
- [13] F. Li, H. Yu, ICD coding from clinical text using multi-filter residual convolutional neural network, *AAAI Conf. Artif. Intel.* 34 (2020) 8180–8187.
- [14] F. Teng, W. Yang, L. Chen, L. Huang, Q. Xu, Explainable prediction of medical codes with knowledge graphs, *Front. Bioeng. Biotechnol.* 8 (2020) 867.
- [15] X. Xie, Y. Xiong, P.S. Yu, Y. Zhu, EHR coding with multi-scale feature attention and structured knowledge graph propagation, in: *ACM International Conference on Information and Knowledge Management*, 2019, pp. 649–658.
- [16] T. Zhou, P. Cao, Y. Chen, K. Liu, J. Zhao, K. Niu, W. Chong, S. Liu, Automatic ICD coding via interactive shared representation networks with self-distillation mechanism, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 5948–5957.
- [17] Z. Yuan, C. Tan, S. Huang, Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding, 2022 arXiv e-prints arXiv:01515.
- [18] H. Schfer, C.M. Friedrich, Umls mapping and word embeddings for ICD code assignment using the MIMIC-III intensive care database, in: *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 6089–6092.

- [19] A.D. Reys, D. Silva, D. Severo, S. Pedro, M.M.d. Sousa e S, G.A. Salgado, Predicting multiple ICD-10 codes from Brazilian-Portuguese clinical notes, in: Brazilian Conference on Intelligent Systems, Springer, 2020, pp. 566–580.
- [20] J. Luo, C. Xiao, L. Glass, J. Sun, F. Ma, Fusion: towards automated ICD coding via feature compression, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2096–2101.
- [21] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780.
- [22] D. Pascual, S. Luck, R. Wattenhofer, Towards Bert-Based Automatic Icd Coding: Limitations and Opportunities, 2021 arXiv e-prints arXiv:2106.06709.
- [23] R. Kaur, J.A. Ginige, Comparative analysis of algorithmic approaches for auto-coding with ICD-10-AM and ACHI, *Stud. Health Technol. Inf.* 252 (2018) 73–79.
- [24] X. Diao, Y. Huo, S. Zhao, J. Yuan, M. Cui, Y. Wang, X. Lian, W. Zhao, Automated icd coding for primary diagnosis via clinically interpretable machine learning, *Int. J. Med. Inf.* 153 (2021), 104543.
- [25] J. Su, Wobert: word-based Chinese BERT model - ZhuiyiAI, Tech. rep. (2020). <https://github.com/ZhuiyiTechnology/WoBERT>.
- [26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A Robustly Optimized BERT Pre-training Approach, 2019 arXiv e-prints arXiv:1907.11692.
- [27] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read Students Learn Better: on the Importance of Pre-training Compact Models, 2019 arXiv e-prints arXiv:1908.08962.
- [28] Z. Zhao, H. Chen, J. Zhang, X. Zhao, T. Liu, W. Lu, X. Chen, H. Deng, Q. Ju, X. Du, Uer: an Open-Source Toolkit for Pre-training Models, 2019 arXiv e-prints arXiv:1909.05658.
- [29] P. Bhargava, A. Drozd, A. Rogers, Generalization in NLI: Ways (Not) to Go beyond Simple Heuristics, 2021 arXiv preprint arXiv:2101.01518.
- [30] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read Students Learn Better: the Impact of Student Initialization on Knowledge Distillation, 2019 arXiv preprint 1908.08962.
- [31] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013 arXiv e-prints arXiv:1301.3781.
- [32] Y. Yu, M. Li, L. Liu, Z. Fei, F.-X. Wu, J. Wang, Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN, *J. Biomed. Inf.* 91 (2019), 103114.
- [33] Y. Kim, Convolutional Neural Networks for Sentence Classification, 2014 arXiv e-prints arXiv:1408.5882.
- [34] J. Han, J. Pei, M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, New York, 2011.
- [35] J.C. Platt, Sequential minimal optimization: a fast algorithm for training support vector machines, in: Report, Advances in Kernel Methods - Support Vector Learning, 1998.
- [36] J.C. Ferrao, F. Janela, M.D. Oliveira, H.M. Martins, Using structured EHR data and svm to support ICD-9-CM coding, in: IEEE International Conference on Healthcare Informatics, IEEE, 2013, pp. 511–516.
- [37] S. Karimi, X. Dai, H. Hassanzadeh, A. Nguyen, Automatic Diagnosis Coding of Radiology Reports: A Comparison of Deep Learning and Conventional Classification Methods, *BioNLP*, 2017, pp. 328–332.
- [38] P. Cao, C. Yan, X. Fu, Y. Chen, K. Liu, J. Zhao, S. Liu, W. Chong, Clinical-coder: assigning interpretable ICD-10 codes to Chinese clinical notes, in: Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 294–301.
- [39] Z. Shuai, D. Xiaolin, Y. Jing, H. Yanni, C. Meng, W. Yuxin, Z. Wei, Comparison of different feature extraction methods for applicable automated ICD coding, *BMC Med. Inf. Decis. Making* 22 (1) (2022) 1–15.
- [40] A. Sonabend, W. Cai, Y. Ahuja, A. Ananthakrishnan, Z. Xia, S. Yu, C. Hong, Automated ICD coding via unsupervised knowledge integration (UNITE), *Int. J. Med. Inf.* 139 (2020) 104135.
- [41] M. Docherty, S.A. Regnier, G. Capkun, M.-M. Balp, Q. Ye, N. Janssens, A. Tietz, J. Lffler, J. Cai, M.C. Pedrosa, J.M. Schattenberg, Development of a novel machine learning model to predict presence of nonalcoholic steatohepatitis, *J. Am. Med. Inf. Assoc.* 28 (6) (2021) 1235–1241.
- [42] M. Subotin, A.R. Davis, A method for modeling co-occurrence propensity of clinical codes with application to ICD-10-PCS auto-coding, *J. Am. Med. Inf. Assoc.* 23 (5) (2016) 866–871.
- [43] L. Zhou, C. Cheng, D. Ou, H. Huang, Construction of a semi-automatic ICD-10 coding system, *BMC Med. Inf. Decis. Making* 20 (2020) 1–12.