

# Towards Automated ICD Coding Using Deep Learning

Haoran Shi<sup>\*1,2</sup>, Pengtao Xie<sup>1</sup>, Zhiting Hu<sup>1</sup>, Ming Zhang<sup>2</sup>, and Eric P. Xing<sup>1</sup>

<sup>1</sup>Petuum Inc, Pittsburgh, USA

<sup>2</sup>Department of Computer Science and Technology, Peking University, Beijing, China

## ABSTRACT

International Classification of Diseases(ICD) is an authoritative health care classification system of different diseases and conditions for clinical and management purposes. Considering the complicated and dedicated process to assign correct codes to each patient admission based on overall diagnosis, we propose a hierarchical deep learning model with attention mechanism which can automatically assign ICD diagnostic codes given written diagnosis. We utilize character-aware neural language models to generate hidden representations of written diagnosis descriptions and ICD codes, and design an attention mechanism to address the mismatch between the numbers of descriptions and corresponding codes. Our experimental results show the strong potential of automated ICD coding from diagnosis descriptions. Our best model achieves 0.53 and 0.90 of F1 score and area under curve of receiver operating characteristic respectively. The result outperforms those achieved using character-unaware encoding method or without attention mechanism. It indicates that our proposed deep learning model can code automatically in a reasonable way and provide a framework for computer-auxiliary ICD coding.

## Introduction

The International Classification of Diseases (ICD) is a health care classification system maintained by the World Health Organization<sup>1</sup>, which provides a hierarchy of diagnostic codes of diseases, disorders, injuries, signs, symptoms, etc. It is widely used for reporting diseases and health conditions, assisting in medical reimbursement decisions, collecting morbidity and mortality statistics, to name a few.

While ICD codes are important for making clinical and financial decisions, medical coding – which assigns proper ICD codes to a patient admission – is time-consuming, error-prone and expensive. Medical coders review the diagnosis descriptions written by physicians in the form of textual phrases and sentences and (if necessary) other information in the electronic medical record of a clinical episode, then manually attribute the appropriate ICD codes by following the coding guidelines<sup>2</sup>. Several types of errors frequently occur. First, when writing diagnosis descriptions, physicians often utilize abbreviations and synonyms, which causes ambiguity and imprecision when the coders are matching ICD codes to those labels<sup>3</sup>. Second, in many cases, several diagnosis descriptions are closely related and should be combined into a single combination ICD code. However, unexperienced coders may code each disease separately. Such errors are called *unbundling*. Third, the ICD codes are organized in a hierarchical structure where the top-level codes represent generic disease categories and the bottom-level codes represent more specific diseases. A miscoding happens when the coder matches the diagnosis description to an overly generic code instead of a more specific code. The cost incurred by coding errors and the financial investment spent on improving coding quality are estimated to be \$25 billion per year in the US<sup>4,5</sup>.

To reduce coding errors and cost, we aim at building an ICD coding machine which automatically and accurately translates the free-text diagnosis descriptions into ICD codes. To achieve this goal, several technical challenges need to be addressed. First, the diagnosis descriptions written by physicians and the textual descriptions of ICD codes are written in quite different styles even if they refer to the same disease. In particular, the definitions of ICD code are formally and precisely worded, while diagnosis descriptions are usually written in an informal and ungrammatical way, with telegraphic phrases, abbreviations, and typos. Second, as stated earlier, there does not necessarily exist a one-to-one mapping between diagnosis descriptions and ICD codes, and human coders should consider the overall health condition when assigning codes. In many cases, two closely related diagnosis descriptions need to be mapped into a single combination ICD code. On the other hand, physicians may write two health conditions into one diagnosis description which should be mapped onto two ICD codes under such circumstances.

---

\*This work was done when the first author was an intern at Petuum Inc.

## Contributions

We present a deep learning approach to automatically perform ICD coding given the diagnosis descriptions. Specifically, we propose a hierarchical neural network model which is able to capture the latent semantics of ICD definitions and diagnosis descriptions, despite their significant difference in writing style. Attention mechanism is designed to address the mismatch between diagnosis description number and assigned code number. We train the model on 8,066 hospital admissions, tune hyper-parameters on 1,728 admissions, and evaluate the performance on a held-out test set of 1,729 hospital admissions. We demonstrate that our coding machine can accurately assign ICD codes.

## Related work

The accuracy and efficiency of manual ICD coding has always been a concern of clinical practice. KJ O'malley *et al.* has summarized the complete workflow of assigning ICD codes manually<sup>2</sup>, which is a dedicated procedure and is prone to errors. To avoid the massive human labour to code, scientists have proposed some automatic or semi-automatic ICD classification system, especially from narrative clinical notes for better health care practice<sup>6-12</sup>. But the experimental dataset was generally small and domain specific. For example, the shared task involved assigning ICD-9 codes to 1954 radiology records has attracted a lot of attention<sup>10</sup>, and In 2015 Koopman *et al.* propose a classification system for identifying different types of cancers for death certificates based on ICD classification system<sup>12</sup>. Contrary to these experiments, the dataset in our experiment is much larger and contains various domains of clinical practice.

There also has been some trials to assign ICD codes utilizing the document of discharge summary<sup>13-15</sup>. Leah S. Larkey and W. Bruce Croft have trained three statistical classifiers with many human-tuned parameters and an ensemble model to give candidate ICD labels, more focused on principal diagnostic code(the most significant diagnostic code), to each discharge summary document<sup>13</sup>. Besides, Franz *et al.* compares three coding methods given discharge diagnosis, but their object is to assign just one diagnostic code to each diagnosis description<sup>14</sup>. All of them utilize the full-text document of discharge summary, thus suffer from the complicated preprocessing of the noisy text. To build a more practical ICD coding machine, we formulate our coding task as a general multi-label classification problem on diagnosis descriptions, without many parameters to tune or restricting the number of assigned codes for each patient record.

## Methods

### Dataset and preprocessing

We perform the study on the publicly available MIMIC-III dataset<sup>16</sup>, which contains de-identified and comprehensive electronic medical records of 58,976 patient visits in the Beth Israel Deaconess Medical Center from 2001 to 2012. The patient visit record usually has a clinical note called discharge summary, which contains multiple sections of information, such as 'discharge diagnosis', 'past medical history', 'admission medications', and 'chief complaint'. The diagnosis descriptions are usually included in the 'discharge diagnosis' and 'final diagnosis' sections<sup>17</sup>. We use a variety of standard text pre-processing techniques such as regular expression matching and tokenization to turn the noisy and irregular raw note texts in these sections into clean diagnosis descriptions. Each resulting label is a short phrase or a sentence, articulating one disease or condition. Patient visits that contain no extracted diagnosis descriptions are discarded.

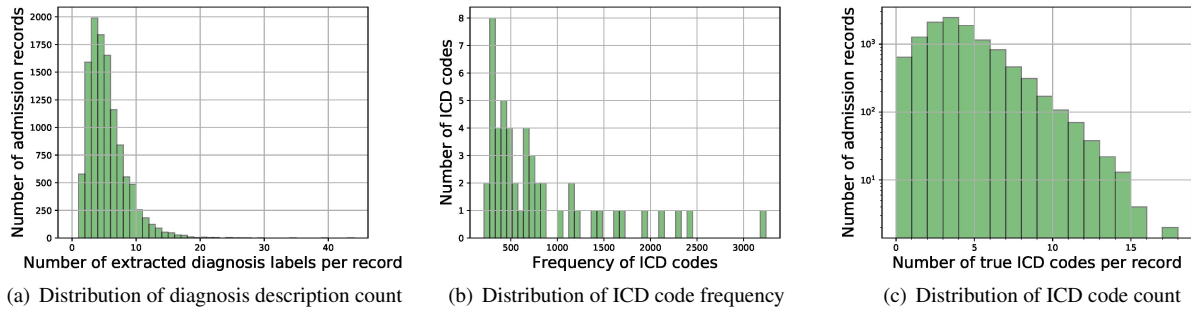
Each patient visit has a list of ICD codes given by the medical coders. These codes are documented in structured tables. The entire dataset contains 6,984 unique codes, each of which has a textual description, describing a disease, symptom, or condition. Many codes are only assigned to a few patient visits. Due to the sparsity of data, it is very difficult to train an accurate coding model for all of them. Instead, we choose 50 most frequent codes to carry out the study while noting that our model can readily be extended to more codes as long as sufficient training data is available. The frequency of one code is measured as the number of patient visits that the code is assigned to.

Table 1 shows a sample of admission record in the raw dataset and extracted diagnosis descriptions. The 'HADMID' is used by the MIMIC-III to denote each hospital admission identically. We omitted irrelevant sections in the original texts of discharge summary, like 'discharge disposition' and 'physical examination'. The extracted diagnosis descriptions given by physicians are in enumeration style. Notice that there is an extra newline in the third written diagnosis description and it's removed after extraction. The number of diagnosis descriptions is not equal to the number of assigned diagnostic codes.

In this way, we obtain 11,523 hospital admission records with overall 59,302 diagnosis descriptions. Figure 1(a) shows the distribution of the number of extracted plain-text diagnosis descriptions across medical records. After restricting our ICD coding target to the 50 most frequent codes, the distribution of ICD code frequency is shown in Figure 1(b), and the distribution of the number of assigned codes per admission record is shown in 1(c). We split the dataset into training set with 8,066 hospital admission records, validation set with 1,728 records, and test set with 1,729 records.

HADMID	189797
Original Texts of Discharge Summary	... DISCHARGE DIAGNOSIS: 1. Prematurity at 35 4/7 weeks gestation 2. Twin number two of twin gestation 3. Respiratory distress secondary to transient tachypnea of the newborn 4. Suspicion for sepsis ruled out ...
Extracted Diagnosis Descriptions	1. Prematurity at 35 4/7 weeks gestation 2. Twin number two of twin gestation 3. Respiratory distress secondary to transient tachypnea of the newborn 4. Suspicion for sepsis ruled out
Assigned ICD Diagnostic Codes	'V3100', '76518', '7756', '7706', 'V290', 'V053'

**Table 1.** One admission sample from MIMIC-III dataset.



**Figure 1.** Distribution of experimental data

## Model design

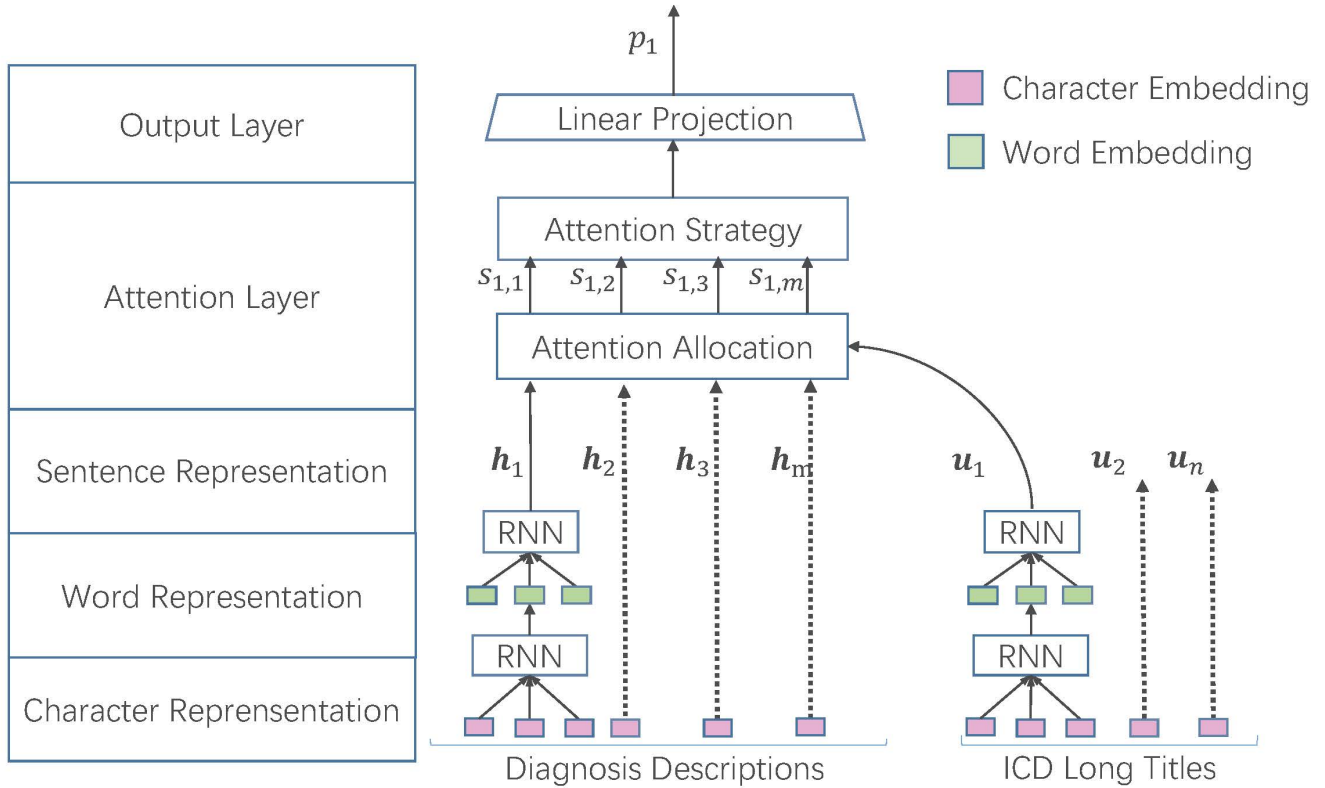
The ICD coding model mainly consists of four modules, which are used for (1) encoding the diagnosis descriptions, (2) encoding the ICD codes based on their textual descriptions, (3) matching diagnosis descriptions with ICD codes, and (4) assigning the ICD codes, respectively. The overall architecture is illustrated in Figure 2. In the following we present each component in detail.

### Diagnosis description encoder

We leverage the long short-term memory (LSTM) recurrent network to encode the diagnosis descriptions<sup>18</sup>. LSTM is a popular variant of the recurrent neural network (RNN). Due to the capacity of capturing long-range semantics in texts, LSTM is widely used for language modeling and sequence encoding<sup>19,20</sup>. An LSTM recurrent network consists of a sequence of units, each of which models one item in the input sequence. Each unit consists of an input gate  $\mathbf{i}$ , a forget gate  $\mathbf{f}$ , a cell gate  $\mathbf{g}$ , an output gate  $\mathbf{o}$ , a cell state  $\mathbf{c}$ , and a hidden state  $\mathbf{h}$ , which are all vectors. They are computed as follows:

$$\begin{aligned}
\mathbf{i}_t &= \text{sigmoid}(\mathbf{W}_{ii}\mathbf{x}_t + \mathbf{b}_{ii} + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_{hi}) \\
\mathbf{f}_t &= \text{sigmoid}(\mathbf{W}_{if}\mathbf{x}_t + \mathbf{b}_{if} + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_{hf}) \\
\mathbf{g}_t &= \tanh(\mathbf{W}_{ig}\mathbf{x}_t + \mathbf{b}_{ig} + \mathbf{W}_{hg}\mathbf{h}_{t-1} + \mathbf{b}_{hg}) \\
\mathbf{o}_t &= \text{sigmoid}(\mathbf{W}_{io}\mathbf{x}_t + \mathbf{b}_{io} + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_{ho}) \\
\mathbf{c}_t &= \mathbf{f}_t * \mathbf{c}_{t-1} + \mathbf{i}_t * \mathbf{g}_t \\
\mathbf{h}_t &= \mathbf{o}_t * \tanh(\mathbf{c}_t)
\end{aligned} \tag{1}$$

For clarity, we denote scalars in plain lowercase letters, vectors in bold lowercase, and matrices in bold uppercase. The operator ‘\*’ in Equation 1 denotes element-wise multiplication and  $t$  represents the time step in the sequence. The sigmoid function is defined as:  $\text{sigmoid}(x) = 1/(1 + \exp(-x))$ , and the tanh function is  $\tanh(x) = (\exp(x) - \exp(-x))/(\exp(x) + \exp(-x))$ .



**Figure 2.** Model Architecture.

For each diagnosis description, we use both character-level LSTM network and word-level LSTM network to obtain its hidden representation. Specifically, in the character-level LSTM,  $\mathbf{x}_t$  is the embedding vector of the  $t^{\text{th}}$  character in the word, and  $T$  is the total number of characters in this word. We select the hidden state of LSTM in the last time step as the hidden representation of the word. In the word-level LSTM,  $\mathbf{x}_t$  is the hidden vector of the  $t^{\text{th}}$  word in the sentence, and  $T$  is the number of words. Similarly, we choose the last hidden state as the representation of the sentence. The reason why we choose character-aware encoding method is there are considerable medical terms with same suffix denoting similar diseases and we expect the character-level LSTM to capture such characteristics. In the following, we denote the hidden representations of the written diagnosis descriptions as  $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_m$ , where  $m$  is the number of extracted diagnosis descriptions in one record.

#### **ICD code encoder**

For each ICD code, we adopt the same two-level LSTM architecture, i.e., character-level and word-level, to obtain the hidden representation of its long title definition, which is provided in the MIMIC-III dataset. For example, in MIMIC-III, the long title of ICD code ‘4010’ is ‘Malignant essential hypertension’. The hidden vector of ‘Malignant essential hypertension’ obtained with the LSTM network serves as the representation of ICD code ‘4010’. The parameters of the neural networks for the ICD code encoder and the diagnosis description encoder are not tied, in order to learn different language styles of these two sets of texts. we use  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$  to denote the hidden representations of different ICD codes obtained by their long title definitions, where  $n$  is the total number of ICD categories. As in our experiment we have picked out the most frequent 50 codes,  $n = 50$ .

#### **Attentional match**

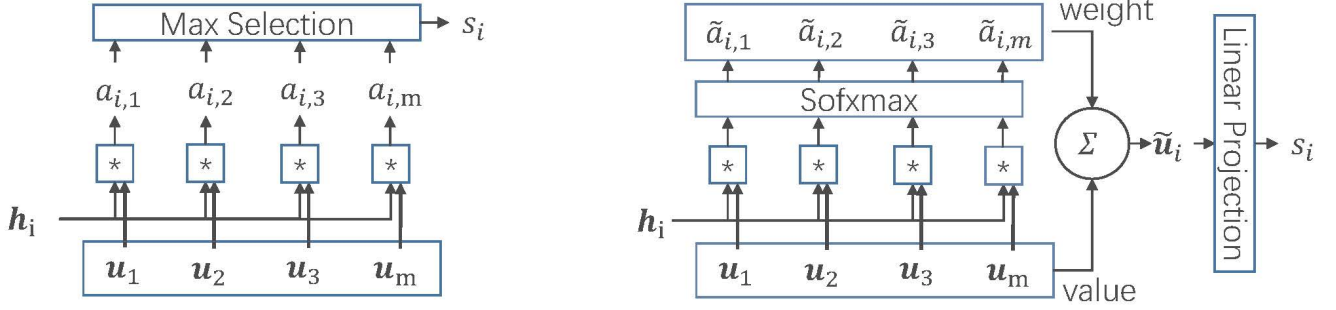
Typically, the number of written diagnosis descriptions does not equal to the number of assigned ICD codes, so we cannot directly assign one code to one diagnosis description. Considering that human coders are supposed to assign appropriate codes according to overall health condition, in parallel, we take all diagnosis descriptions into account during coding by adopting an attention strategy. The attention mechanism provides a recipe for choosing which diagnosis descriptions are important when performing coding.

We use  $u_{i,k}$  and  $h_{j,k}$  to represent the  $k^{\text{th}}$  dimension of hidden representations of the  $i^{\text{th}}$  ICD code and the  $j^{\text{th}}$  diagnosis description, respectively. For the  $i^{\text{th}}$  ICD code, we use  $a_{i,j}$  to denote its attention score on the  $j^{\text{th}}$  diagnosis description, which is

the cosine similarity of the hidden representations of the  $i^{th}$  ICD code and the  $j^{th}$  diagnosis description.

$$a_{i,j} = \sum_{k=1}^d u_{i,k} h_{j,k} \quad (2)$$

Then we design two different kinds of attention layers to obtain the confidence score of ICD code assignment: *Hard-selection* and *Soft-attention* mechanism, which are depicted in Figure 3.



**Figure 3.** Two Architectures of Attentional layer in our model. The ‘\*’ in the square denotes inner-product calculation. **Left** is the Hard-selection Mechanism. **Right** is the Soft-attention Mechanism.

**Hard-selection.** Based on the assumption that the most related diagnosis description plays a decisive role when assigning ICD code, for each ICD code, we define the dominating diagnosis as the one that has the maximum attention score among all diagnosis descriptions. We apply the sigmoid function to normalize the score into a probability value in  $[0, 1]$ . The probability of the  $i^{th}$  ICD code being assigned is thus:

$$p_i = \text{sigmoid}(\max_{j=1,2,\dots,m} (a_{i,j})) \quad (3)$$

**Soft-attention.** Instead of choosing the single maximum attention score, here we apply a softmax function to normalize the attention scores among all diagnosis descriptions into a probability simplex. The normalized attention scores are utilized as the weights of different diagnosis descriptions. We then use the weighted average over the hidden representations of different diagnosis descriptions as the attentional hidden vector. In this way, the attentional hidden vector can take into account all diagnosis descriptions with varying levels of attention. The attentional vector of the  $i^{th}$  ICD code is denoted as  $\tilde{u}_i$ .

$$\tilde{a}_{i,j} = \frac{\exp(a_{i,j})}{\sum_{j=1}^m (\exp(a_{i,j}))} \quad (4)$$

$$\tilde{u}_i = \sum_{j=1}^m \tilde{a}_{i,j} * h_j \quad (5)$$

### Linear projection layer

For the attentional hidden vector  $\tilde{u}_i$ , we use linear Perceptron structure as the output layer to project the vector into a real value<sup>21,22</sup>, which represents the confidence score of predicting label to be true. The Perceptron parameters are different among each code. Finally, we utilize sigmoid function to normalize the confidence score into a probability, which ranges from 0 to 1.

$$s_i = \sum_{k=1}^d w_{i,k} \tilde{u}_{i,k} \quad (6)$$

$$p_i = \text{sigmoid}(s_i) \quad (7)$$

### Parameter learning

We use binary cross entropy as the loss function for each ICD code<sup>23,24</sup>. The loss function for each pat record can be formulated as follows:

$$Loss = -1/n \sum_{j=1}^n (t_i * \log(p_i) + (1 - t_i) * \log(1 - p_i)), \quad (8)$$

where  $t_i$  is the real label of the  $i^{th}$  ICD code, i.e., true(1) or false(0). All parameters are learned by minimizing the loss function with stochastic gradient descent<sup>25</sup>.

### Hyperparameter setting

The model is trained on the training set using the standard ADAM optimizer<sup>26</sup>, with an initial learning rate 0.001 and mini-batch size 10. Hyper-parameters are fine-tuned on the validation set. In particular, the number of hidden units and output units of all LSTM modules are 200. For word-level LSTM in our experiment, we apply a dropout layer with 0.5 dropout probability to the output, to avoid the overfitting problem<sup>27</sup>. Since our model provides a probability score for each assignment of ICD code, we also tune on the validation set the optimal threshold that cuts the probability score into a binary output, i.e., true or false, to obtain best F1 score.

### Analysis and evaluation

Considering the ICD code assignment is generally sparse, with most ICD codes labeled as false and only a few as true, we use the micro F1-score and AUC\_ROC (area under curve of receiver operating characteristic) score as the quantitative metrics. Micro F1-score is a harmonic mean of precision and recall. It is widely used to evaluate the performance of a binary classifier on imbalanced data<sup>28</sup>. The micro AUC\_ROC score is calculated as the area under the ROC curve, which is drawn by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings<sup>29,30</sup>. Intuitively, the AUC\_ROC score measures the probability that the model assigns higher score for a positive instance than negative one, the lower bound of which is 0.5.

## Results

Table 2 shows the F1-score and AUC\_ROC score of different models evaluated on the test set. We can observe that with the Soft-attention mechanism, the F1 and AUC\_ROC increase 5.2 and 2.3 percent, respectively, compared to the Hard-selection model. To further explore the efficacy of different modules in our model, we perform an ablation study on our intact Soft-attention model, which leverages character-level and word-level LSTM encoder and Soft-attention mechanism. Performance decreasing on several ablation experiments indicates that all the designed modules in our model are necessary and play a crucial role in the coding process. The attention scores on the subset of 50 ICD codes is shown in Table 3 for one patient visit sample. We can

Model Architecture	F1	AUC_ROC
Hard-selection Model	0.480	0.877
Soft-attention Model	<b>0.532</b>	<b>0.900</b>

Ablation Study on Soft-attention Model		
Replace character-level LSTM with random initialized and tunable word embedding	0.508	0.882
Replace character-level LSTM with pre-trained and tunable word embedding	0.528	0.895
Replace word-level LSTM encoder with average encoder	0.504	0.886
Replace attention mechanism with naïve linear classifier	0.471	0.882

**Table 2.** Performance on different models

observe that for different ICD codes, our model allocates different attention scores to diagnosis descriptions automatically. For example, when assigning the ICD code titled ‘Neonatal jaundice associated with preterm delivery’, the model puts more attention on ‘Prematurity at 34 and 5/7 weeks gestation’ and ‘Hyperbilirubinemia of prematurity’, while less attention on other irrelevant diagnosis descriptions like ‘Sepsis ruled out’. The attention allocation results of all 50 the ICD codes for this sample can be found in the supplementary materials.

## Discussion

Our model achieves attractive performance on the ICD coding task, which indicates our architecture is reasonable: to extract diagnosis descriptions from discharge summary, and use the hidden meaning of these descriptions to predict on target ICD code whose meaning is represented by its formal title, can be a promising methodology to perform automated ICD coding. The Soft-attention mechanism can push the model to allocate different attention on multiple diagnosis descriptions, which can obtain better performance compared to Hard-selection. In the following, we will discuss several important insights provided by our ablation study on the Soft-attention model in detail.

Index	Diagnosis description
1	Prematurity at 34 and 5/7 weeks gestation
2	Twin # 2
3	Status post transitional respiratory distress
4	Sepsis ruled out
5	Hyperbilirubinemia of prematurity

LONG TITLE OF ICD CODE	1	2	3	4	5
Esophageal reflux	0.20	0.09	0.12	<b>0.52</b>	0.07
Acute kidney failure with lesion of tubular necrosis	<b>0.48</b>	0.14	0.03	0.08	0.27
Acute kidney failure, unspecified	0.43	0.11	0.04	0.11	0.31
Chronic kidney disease, unspecified	<b>0.44</b>	0.21	0.02	0.18	0.16
Urinary tract infection, site not specified	0.24	0.06	0.01	<b>0.65</b>	0.04
<b>Neonatal jaundice associated with preterm delivery</b>	<b>0.57</b>	0.14	0.00	0.00	0.28
Septic shock	<b>0.58</b>	0.21	0.00	0.12	0.09
Severe sepsis	<b>0.57</b>	0.16	0.00	0.16	0.09
Cardiac complications, not elsewhere classified	0.35	<b>0.41</b>	0.00	0.13	0.11
Aortocoronary bypass status	0.37	<b>0.38</b>	0.02	0.10	0.14
Percutaneous transluminal coronary angioplasty status	0.31	<b>0.45</b>	0.02	0.07	0.14
Long-term (current) use of anticoagulants	0.23	<b>0.51</b>	0.02	0.11	0.13
Long-term (current) use of insulin	0.24	<b>0.31</b>	0.05	0.25	0.15
<b>Observation for suspected infectious condition</b>	0.16	0.10	0.10	<b>0.56</b>	0.08
Single liveborn, born in hospital, delivered without mention of cesarean section	0.02	<b>0.96</b>	0.00	0.01	0.01
Single liveborn, born in hospital, delivered by cesarean section	0.02	<b>0.95</b>	0.00	0.01	0.02
<b>Need for prophylactic vaccination and inoculation against viral hepatitis</b>	0.34	<b>0.38</b>	0.02	0.10	0.16
Personal history of tobacco use	<b>0.52</b>	0.19	0.01	0.17	0.11

**Table 3.** Attention allocation on one sample. **Top** shows all the extracted diagnosis descriptions from written discharge diagnosis given by physicians. **Down** shows the attention allocation on a subset of ICD codes. The codes in bold format have true labels.



To evaluate the effectiveness of the character-level LSTM module in learning the hidden representation of medical vocabulary, we remove it from the model. Instead, we obtain the hidden vector of each word from a tunable word embedding layer, where each word is assigned a fixed-dimension vector. Note that the other modules remain intact and the Soft-attention strategy is leveraged. It causes 0.024 and 0.018 drop of the F1 score and AUC\_ROC respectively, which demonstrates the necessity of incorporating character-level encoder into representation learning. We have also tried to initialize the word embedding layer with pre-trained word vectors. These vectors were learned using word2vec tool on a large corpus of medical research papers collected from Pubmed, BioMed and PLOS<sup>31</sup>. In this setting, All the words are transformed to lowercase and lemmatized in advance. The performance has increased compared to randomly initialized word embeddings but it's still lower than our character-level LSTM model.

To ensure our character-level LSTM module can give reasonable representations for diagnosis description, we have checked the nearest neighbors of words and sentences based on Euclidean distance in hidden space. Table 4 shows a subset of words and sentences and their nearest neighbors. On top of the table displays the word neighbor relationship contrast between the model with character-level LSTM and word embedding layer with pre-trained word vectors. First, it indicates that character-level LSTM word encoder can correct various typos and recognize different morphologies appearing in the written diagnosis descriptions, by generating similar representations for them. For example, our model can recognize different written variants of 'Ischemia' and generate near representations for them. In addition, many disease names and procedures with same suffix are denoting similar diseases, which can be captured by our character-level LSTM encoder efficiently. Otherwise, there exist some words with same suffix but unrelated meanings indeed that are also distributed near in the hidden space, however, these words are not denoting disease categories in many cases, like 'state', so it should have little effect to the coding. While looking at the sentence neighbor relationship shown at the bottom of the table 4, we could observe that our model can generate near embeddings for similar sentences too.

Besides word-level LSTM encoder, word averaging method also shows strong performance to generate sentence embeddings<sup>32</sup>. we have also tried averaging the word embeddings in one sentence to obtain the hidden vector of sentence, instead of using LSTM encoder. Keeping other modules intact, the F1 drops 0.028 and AUC\_ROC drops 0.014, which indicates the word-level LSTM encoder is superior to word averaging.

We evaluate the necessity of the attention mechanism by comparing our Soft-attention model with a linear classifier without attention mechanism. We design the architecture of linear classifier as follows. For each ICD code, we concatenate its hidden vector with the representation of the overall diagnosis, which is obtained by averaging the hidden vectors of each diagnosis description. The concatenated vector is processed by a linear Perceptron to get the confidence score. The parameters of linear Perceptron are independent among different ICD codes. Replacing attention mechanism with such a linear classifier causes the F1 and AUC\_ROC to drop 0.061 and 0.018 respectively, which demonstrates the advantage of our attention mechanism.

## Limitations

The performance achieved by our hierarchical neural models with Soft-attention mechanism shows that reliable performance could be obtained even through a simple diagnosis description extraction process. But, considering the noisy format of the electrical discharge summary, with a more elaborate diagnosis extraction preprocessing and cleaner corpus with high-quality diagnosis descriptions, we believe the performance could be improved further.

Another limitation of our study is the candidate ICD codes are restricted to the most frequent 50 ones. If one ICD code is too rare, there will not be enough evidence to construct a valid neural model, and the label imbalance problem will be more severe<sup>33</sup>, which makes the learning harder. It should be helpful to obtain more formatted records to support the model to learn.

With separate linear Perceptrons to assign each ICD code, we have assumed that the assignment of different ICD codes are mutually independent. However, ICD codes indeed correlate with each other to some extent. For example, the long title name of ICD code '40390' is 'Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified', while '40391' is 'Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage V or end stage renal disease'. If ICD code '40390' is assigned to one patient record, ICD '40391' should definitely not be assigned since these two codes represent exclusive health conditions. And it might be helpful to leverage the hierarchy structure of ICD codes<sup>15</sup>. Thus, modeling such correlations with some structured methods can be meaningful for improving performance.

## Conclusions

We find it is promising to construct a high quality ICD coding machine directly from diagnosis description in the electronic medical records. Our model achieves high performance, suggesting that an attentional match between the diagnosis descriptions and the textual definition of ICD code suits well for the inference task. The proposed Soft-attention mechanism can learn to allocate varying attention strengths on multiple diagnosis descriptions when assigning ICD codes. Just like the reasoning of



Word encoding method	Neighbors	
Character-level LSTM	Ischemia	<i>Ischmia</i> (2.76), ischemia (4.05), Dysthymia (5.46), hypercalcemia (5.74), <i>ishemia</i> (5.76)
	Pneumonia	<i>Penumonia</i> (0.37), <i>Pnuemonia</i> (0.85), <i>pnuemonia</i> (1.77), <i>Pnemonia</i> (1.78), pneumonia (1.95)
	Gastroenteritis	gastroenteritis (2.59), Osteoarthritis (3.99), endomyometritis (4.18), Gastritis (4.18), prostatitis (4.28)
	Coronary	coronary (2.04), Rotary (4.47), ovary (6.60), Artery (7.16), aortopulmonary (7.16)
	State	state (4.04), resuscitate (4.73), prostate (4.81), lactate (4.99), Methotrexate (5.06)
Word2vec	ischemia	ischemic (12.51), reperfusion (15.68), malperfusion (17.87), infarction (18.10), Microinfarction (18.39)
	pneumonia	tracheobronchitis (16.02), tracheitis (16.32), epiglottitis (16.84), abcess (17.07), septicemia (17.09)
	gastroenteritis	diarrhea (17.43), tracheitis (17.90), enteritis (17.91), stools (17.98), diarrheal (18.10)
	coronary	multivessel (16.80), vessels (16.87), atherectomy (16.96), aortoiliac (17.15), pectoris (17.32)
	state	the (15.44), and (15.66), a (15.67), transition (15.71), pauses (15.92), of/Of (16.12)
	<i>ishemia</i>	4.3 (8.33), t4-t5 (8.39), d'or (8.40), bronchiectesis (8.44), difficile (8.44)
	<i>pnuemonia</i>	10% (8.28), 76 (8.35), enterovaginal (8.36), penicillion (8.39), secudnum (8.40)

Diagnosis descriptions	Neighbors
Coronary artery disease	Coronary Artery Disease (0.78), Acute coronary artery disease (1.36), Coronary artery disease stable (1.75), Chronic coronary artery disease (1.82), History of coronary artery disease (1.88)
Congestive heart failure	congestive heart failure (1.66), Pulmonary hypertension / congestive heart failure (1.76), Diastolic congestive heart failure (1.96), Biventricular congestive heart failure (1.99), Diastolic Dysfunction Heart Failure (2.04)
Hemodynamic monitoring with central venous catheter	Bladder atony status post suprapubic catheter (2.33), Patient has suprapubic catheter (2.48), hepatic dysfunction which is resolving (2.70), Erosions in the stomach and duodenum (2.72), ? Gastrointestinal bleed (2.76)
Apnea of prematurity	Apnea of Prematurity (0.27), Apnea of prematurity ongoing (1.41), Retinopathy of prematurity (1.45), Apnea and bradycardia of prematurity (1.47), Apnea bradycardia of prematurity (1.56)
Diabetes mellitus type 2	Diabetes mellitus , type 2 (0.35), Diabetes mellitus Type 2 (0.52), Diabetes mellitus , Type 2 (0.73), Diabetes mellitus 2 (1.17), Diabetes Type 2 (1.20)

**Table 4.** Hidden vector location relationship. **Top** is a comparison between the character-level LSTM encoder and the word2vec model with pre-trained word vectors. The italic text means typo appearing in the diagnosis descriptions. **Bottom** is the diagnosis neighbor relation in our intact Soft-attention model.

human, with more attention on informative diagnosis descriptions and less on irrelevant ones, the Soft-attention model can assign codes based on diagnosis descriptions automatically and efficiently.

Our experiment indicates the potential for real life applications in view of the high performance even on some noisy-formatted data. We believe that with more elaborate data preprocessing techniques, and with more formatted electrical medical records, the automatic coding can be even more accurate. Since our model can give a probability score when assigning ICD code, we can decrease the probability threshold to get higher recall rate or increase to get higher precision. Thus, in addition to coding ICD diagnosis directly, our model can also serve as an assistant tool for doctors, helping them to pre-select a small set of candidate codes and thus greatly alleviating doctors' workloads.

In this paper we have adopted ICD-9 diagnostic codes as coding target, however the proposed approach can straightforwardly be adapted to new revisions of ICD codes, like ICD-10<sup>34</sup>, as long as the formal definitions of all codes and golden diagnostic codes on training data are available.

## References

1. Organization, W. H. *et al.* International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index. *World Heal. Organ.* (1978).
2. O'malley, K. J. *et al.* Measuring diagnoses: Icd code accuracy. *Heal. services research* **40**, 1620–1639 (2005).
3. Sheppard, J. E., Weidner, L. C., Zakai, S., Fountain-Polley, S. & Williams, J. Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping. *Arch. disease childhood* **93**, 204–206 (2008).
4. Lang, D. Consultant report-natural language processing in the health care industry. *Cincinnati Child. Hosp. Med. Center, Winter* (2007).
5. Farkas, R. & Szarvas, G. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics* **9**, S10 (2008).
6. Spyns, P. Natural language processing. *Methods information medicine* **35**, 285–301 (1996).
7. Hearst, M. A. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 3–10 (Association for Computational Linguistics, 1999).
8. Zeng, Q. T. *et al.* Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics decision making* **6**, 30 (2006).
9. Ananiadou, S. & McNaught, J. *Text mining for biology and biomedicine* (Artech House London, 2006).
10. Pestian, J. P. *et al.* A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, 97–104 (Association for Computational Linguistics, 2007).
11. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., Hurdle, J. F. *et al.* Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inf.* **35**, 44 (2008).
12. Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A. & Grayson, N. Automatic icd-10 classification of cancers from free-text death certificates. *Int. journal medical informatics* **84**, 956–965 (2015).
13. Larkey, L. S. & Croft, W. B. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 289–297 (ACM, 1996).
14. Franz, P., Zaiss, A., Schulz, S., Hahn, U. & Klar, R. Automated coding of diagnoses—three methods compared. In *Proceedings of the AMIA Symposium*, 250 (American Medical Informatics Association, 2000).
15. Perotte, A. *et al.* Diagnosis code assignment: models and evaluation metrics. *J. Am. Med. Informatics Assoc.* **21**, 231–237 (2013).
16. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. data* **3** (2016).
17. Prakash, A. *et al.* Condensed memory networks for clinical diagnostic inferencing. In *AAAI*, 3274–3280 (2017).
18. Sundermeyer, M., Schlüter, R. & Ney, H. LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association* (2012).
19. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J. & Khudanpur, S. Recurrent neural network based language model. In *Interspeech*, vol. 2, 3 (2010).
20. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

21. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. review* **65**, 386 (1958).
22. Widrow, B. & Lehr, M. A. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proc. IEEE* **78**, 1415–1442 (1990).
23. Deng, L.-Y. The cross-entropy method: a unified approach to combinatorial optimization, monte-carlo simulation, and machine learning (2006).
24. Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).
25. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, 177–186 (Springer, 2010).
26. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
27. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. machine learning research* **15**, 1929–1958 (2014).
28. Van Rijsbergen, C. Information retrieval. dept. of computer science, university of glasgow. URL: [citeseer.ist.psu.edu/vanrijsbergen79information.html](http://citeseer.ist.psu.edu/vanrijsbergen79information.html) **14** (1979).
29. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiol.* **143**, 29–36 (1982).
30. Fawcett, T. An introduction to roc analysis. *Pattern recognition letters* **27**, 861–874 (2006).
31. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A neural probabilistic language model. *J. machine learning research* **3**, 1137–1155 (2003).
32. Wieting, J., Bansal, M., Gimpel, K. & Livescu, K. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198* (2015).
33. Japkowicz, N. & Stephen, S. The class imbalance problem: A systematic study. *Intell. data analysis* **6**, 429–449 (2002).
34. Organization, W. H. *et al.* International statistical classification of diseases and health related problems, 10th revision. Geneva: WHO (1992).

## Acknowledgements

The authors thank Devendra Singh Sachan for his sharing of pre-trained word vectors.

## Author contributions statement

H.S. and P.X. conceived and designed the study. H.S. processed the data and performed the experiments. H.S., P.X., Z.H. wrote the paper. M.Z. and E.P.X. take responsibility for the paper as co-senior authors. All authors reviewed the manuscript.

## Additional information

### Supplementary information

The supplementary material is available on request.

### Competing interests

The authors declare that they have no competing interests.