Intelligent Medicine 2 (2022) 161–173

Contents lists available at ScienceDirect

Intelligent Medicine

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

Review

A survey of automated International Classification of Diseases coding: development, challenges, and applications

Chenwei Yan[1,2], Xiangling Fu[1,2,*], Xien Liu[3,*], Yuanqiu Zhang[1,2], Yue Gao[1,2], Ji Wu[3], Qiang Li[4]

[1] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing 100876, China
[3] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[4] Beijing Tsinghua Changgung Hospital, Beijing 102218, China

ARTICLE INFO

Keywords:
International Classification of Diseases coding
Disease classification
Health-related document
Electronic medical record
Medical record management
Clinical coding

ABSTRACT

The International Classification of Diseases (ICD) is an international standard and tool for epidemiological investigation, health management, and clinical diagnosis with a fundamental role in intelligent medical care. The assignment of ICD codes to health-related documents has become a focus of academic research, and numerous studies have developed the process of ICD coding from manual to automated work. In this survey, we review the developmental history of this task in recent decades in depth, from the rules-based stage, through the traditional machine learning stage, to the neural-network-based stage. Various methods have been introduced to solve this problem by using different techniques, and we report a performance comparison of different methods on the publicly available Medical Information Mart for Intensive Care dataset. Next, we summarize four major challenges of this task: (1) the large label space, (2) the unbalanced label distribution, (3) the long text of documents, and (4) the interpretability of coding. Various solutions that have been proposed to solve these problems are analyzed. Further, we discuss the applications of ICD coding, from mortality statistics to payments based on disease-related groups and hospital performance management. In addition, we discuss different ways of considering and evaluating this task, and how it has been transformed into a learnable problem. We also provide details of the commonly used datasets. Overall, this survey aims to provide a reference and possible prospective directions for follow-up research work.

1. Introduction

The International Classification of Diseases (ICD) is an internationally unified disease classification method developed by the World Health Organization. It classifies diseases according to their etiology, pathology, clinical manifestations, and anatomical location in a systematic fashion, using coding to represent diseases. Assigning the appropriate ICD codes to health-related documents is called ICD coding. Using this unified coding of diseases is conducive to storage, retrieval, and analysis of medical data. ICD is also an effective means of standardizing medical data, as well as forming the data basis for intelligent medicine applications.

ICD taxonomy is also the basis of the diagnosis-related groups (DRGs) payment system. Wrong coding, omissions, or multiple coding will directly affect the quality of the data used for DRGs; the loss caused by such errors in ICD code assignment cannot be underestimated. According to statistics from the US Centers for Medicare and Medicaid, the error payout rate in 2000 was 6.8%; these errors were mainly caused by incorrect ICD code assignment [1]. Another survey report indicated that the cost of correcting wrong codes of ICD-9-CM could reach US $25 billion per year [2]. Thus, coding errors not only have an impact on patients' follow-up treatment but also cause significant losses to the medical expense payment system. Therefore, from a practical perspective, ICD is of great significance as a means of improving the efficiency of medical record management, reducing the cost of DRGs, and minimizing loss in medical payments.

In the early stages, ICD coding relied on manual work by professional coders. After browsing the available documentation, such as electronic medical records (EMRs), the professional coder would manually select one or multiple suitable ICD codes from the tens of thousands of codes that exist and assign them to the medical record. This process is error-prone and time-consuming. Moreover, it requires coders to have good medical knowledge and to be familiar with coding specifications and rules. It is also costly, as professional coders need to be trained regularly to keep up with the continuous updates of the ICD. From ICD-9-CM to ICD-10-CM, the number of ICD codes increased more than five-fold [3].

* Corresponding authors: Xiangling Fu, School of Computer Science (National Demonstrative Software School), Beijing University of Posts and Telecommunications, Beijing 100876, China (Email: fuxiangling@bupt.edu.cn); Xien Liu, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (Email: xeliu@mail.tsinghua.edu.cn).

https://doi.org/10.1016/j.imed.2022.03.003
Received 15 November 2021; Received in revised form 5 January 2022; Accepted 11 March 2022

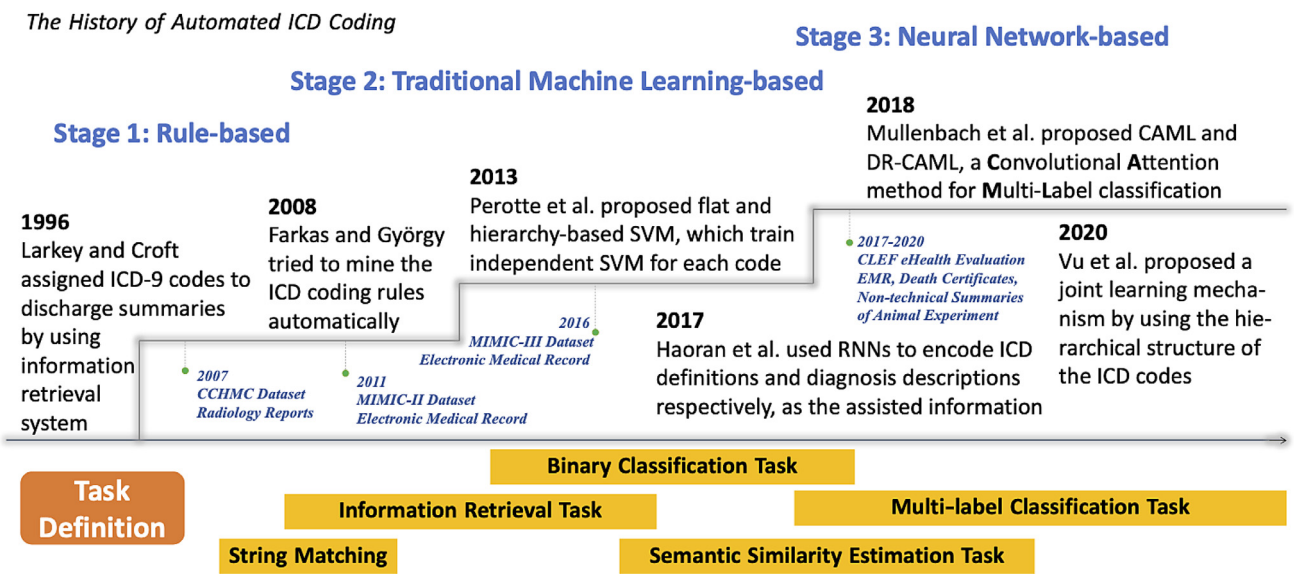

Figure 1. Developmental history of the automated ICD coding task.

The newest, and the latter version of the ICD codes can classify patients' conditions, injuries, and diseases in more detail than previous versions, resulting in substantial increases in both the number of codes and the difficulty of manual coding.

To this end, researchers have proposed various automated coding approaches. The earliest research on this topic [4–5], to the best of our knowledge, took place in the 1990s. We review in detail the progress made in this task in Section 4. Figure 1 provides an overview of the history of automated ICD coding research, showing the three development stages, milestones such as critical research or data publications, and how our understanding of the task has changed. Detailed task definitions will be introduced in Section 3.

Despite many advances in academic research, this task still faces severe challenges in clinical applications. The first challenge is the large label space. Regarding a code as a label, the number of labels can reach tens of thousands. This huge search space makes both retrieval and classification very difficult. The second challenge is the extremely unbalanced label distribution [6]. A few ICD codes occur frequently in health-related documents, whereas most codes are uncommon. The third challenge is the long document representation. The key is to extract useful text segments to assist a models to map the diagnosis to ICD codes. Last but not least, the interpretability of coding represents a fourth challenge. It is essential to provide a reasonable explanation for predicted codes in actual applications. These challenges limit the practical applications of coding tasks in clinical practice.

2. Characteristics of ICD codes

The ICD taxonomy organizes codes in a hierarchical architecture consisting of chapters, sections, and categories, even expansion codes. Figure 2 illustrates an example of *Escherichia coli* (*E. coli*) infection in the ICD-11-MMS system. Assuming that every ICD code is a node, the section codes and their corresponding chapter codes can be treated as child–parent relations. Thus, a node at one level may be the parent of a node at the next level. Nodes that have the same parent node can be regarded as sibling nodes. Clearly, child nodes are semantically similar to their parents. It can be inferred that child nodes inherit some information from their parent node, whereas sibling nodes at the same level represent refinements and extensions of the parent node with respect to different aspects. Therefore, the difference between sibling nodes tends to be mutually exclusive. Taking the *E. coli* infection in Figure 2 as an example, "1A03.1" and "1A03.2" are sibling nodes, and both belong to the diagnosis of "*E. coli* infection". One is "Enterotoxigenic" and the other is "Intestinal invasive." Thus, they are unlikely to be assigned to the same EMR at the same time. Furthermore, certain codes under different parent nodes often appear in the same EMR owing to the complicated internal logical relationship behind the disease (such as cold and cough); we call these nodes friend nodes. They seems to have no connection in the tree-like structure, but they often appear in pairs in EMRs.

Overall, these characteristics of parent–child nodes, sibling nodes, and friend nodes in the ICD taxonomy can be classified into categories of inheritance, mutual exclusion, and co-occurrence. A detailed analysis is presented in the following parts.

2.1. Parent-child nodes: inheritance

ICD codes are organized in a hierarchical structure. The parent codes are often a general disease category, and the child codes are refinements and supplements of their parent codes with respect to a certain disease feature. For example, the category code "E11" in ICD-10 represents type 2 diabetes; its child code "E11.3" means type 2 diabetes with ocular complications, which is a supplement to the parent code in terms of complications; and the code "E11.302" represents type 2 diabetic cataract, which is a specific disease of ocular complications. Therefore, although we are only required to predict a specific disease when we predict an ICD code assign child ICD codes when we predict concrete diseases, the parent node also contains very relevant and useful information that can give an approximate direction.

2.2. Sibling nodes: mutual exclusion

Sibling nodes in the same layer of the ICD taxonomy are classified by their axis, which includes etiology, pathology, anatomical location, and clinical manifestations. Most categories and sub-categories have only one axis; there are two classification axes in a few cases. This means that for a given layer, codes divided classified by the same axis are often mutually exclusive. According to a survey [7], the third-ranking cause of coding defects is logical conflicts of the classification axis, accounting for 9.6% of all ICD coding defects. Specific logical conflicts include codes representing a condition with with and without complications being assigned to one medical record, or traumatic and non-traumatic codes for one disease being assigned at the same time.

2.3. Friend nodes: co-occurrence

From the perspectives of disease complications and etiology, many diseases are related, and this relationship may even be causal. Patients

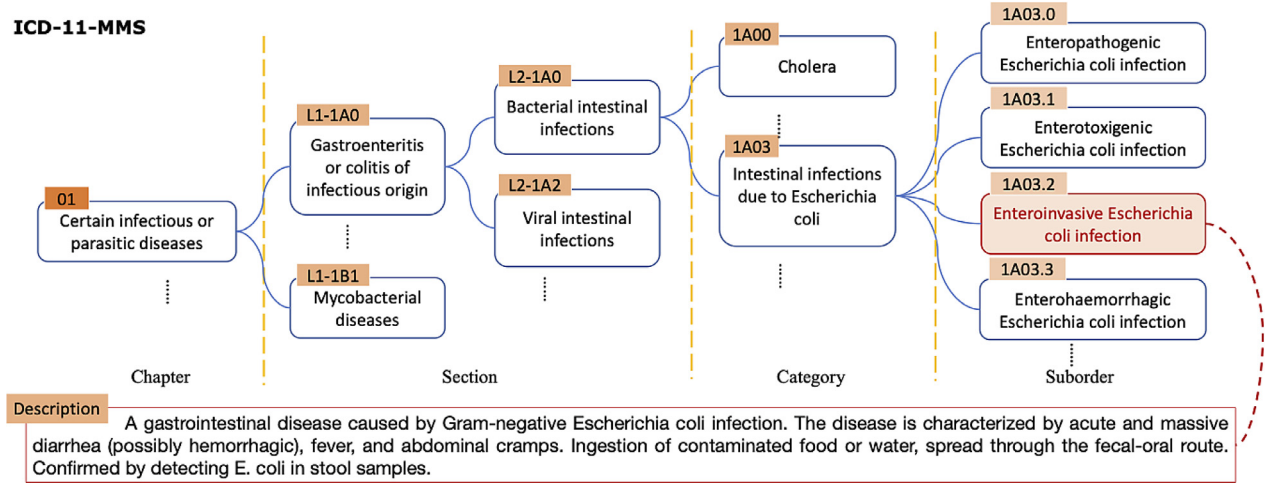

Figure 2. An example of *E. coli* infection in the ICD-11-MMS system.

diagnosed with disease A often suffer from disease B or disease C. For example, in the medical records of patients with acute myocardial infarction, heart failure (a common complication), coronary atherosclerosis (cause), or other related diseases often appear as secondary diagnosis. The typical representative of this phenomenon is co-morbidity recognition [8–10]. The correlations between types of disease are also reflected in the process of ICD coding; this is the co-occurrence aspect of coding.

These three relations between ICD codes constitute the basic characteristics of ICD taxonomy. They enable the codes to be relevant and meaningful, that is, they are no longer independent labels. This distinguishes the ICD coding task from general multi-label tasks.

3. Task definition and evaluation metrics

3.1. Task definition

From the 1990s to the present, as our understanding of ICD coding has evolved, this problem has been treated as different types of task. Thus, we first summarize how researchers view the task of ICD coding and how they transform it into a learnable problem from different perspectives.

Information Retrieval. Initially, most studies treated automated ICD coding as an information retrieval task, using string matching, statistical processing, part-of-speech tagging, negative recognition, and even medical language processing tools such as MedLEE to identify medical concepts from structured and unstructured text [11], so that the diagnosis could be matched with ICD codes [12–14]. Most documents used in these studies were from EMRs.

Binary Classification. With the rise of machine learning, many scholars began to regard the problem as a binary classification task. They trained an independent classifier for each ICD code to determine whether a code should be assigned to the corresponding document, and then integrated the prediction results of all the classifiers as the final prediction result [15-16]. This classification method is often used in situations where there are not many ICD codes, such as radiology reports.

Semantic Similarity Estimation. In addition, some studies [17-18] have defined ICD coding as a semantic similarity estimation task. Such approaches only utilize the short text describing the actual diagnosis (the remainder of the document is discarded) and compare the diagnosis with the disease description in the ICD coding taxonomy [19]. The common computation way between text and label is cosine similarity [20]. Such methods require high accuracy and standardization of the doctor's written diagnosis.

Machine Translation. Moreover, the task may be viewed as a special type of machine translation problem [21]. It is assumed that the input (diagnostic terms) and the output (codes) are two different languages, and that sequence-to-sequence models are suitable for the translation from one to another. Similar to semantic similarity estimation, only the short text of the diagnosis is used as the input, rather than the whole document.

Multimodal Machine Learning. Benefiting from the abundance of data, some studies [11,16,22] have made full use of structured and semi-structured data, such as demographic data or laboratory results. Such methods are often limited by the availability of datasets, but they represent a promising direction.

Multi-label Classification. In recent years, automated ICD coding has increasingly often been treated as a multi-label classification task [23–25]. The whole document is used as the input to obtain a text representation, each ICD code is regarded as a label, and all the labels are predicted simultaneously. The document type also extends to death certifications [26], non-technical summaries (NTS) [27], etc.

There are two distinct differences in ICD coding multi-class text classification compared with general multi-class text classification. First, the labels of the automated ICD coding task are organized in a hierarchical structure. As summarized in Section 2, ICD codes exhibit the characteristics of inheritance, mutual exclusion, and co-occurrence. Second, as medical text is knowledge-intensive, text classification models used in the general domain may not be able to understand the medical semantics used in health-related documents.

3.2. Evaluation metrics

Many evaluation metrics have been proposed to evaluate the effectiveness of the methods from different perspectives. The main evaluation metrics are precision, recall, and F1, and each is calculated by micro-average and macro-average approaches.

The area under the receiver operating characteristic curve (AUC) and $P@k$ (precision at k, the fraction of the k highest-scoring labels that are present in the ground truth) are also commonly used. For decision support, it is convenient to show a fixed number of predictive codes to users. The value of k is usually selected as 5, 8, or 15 [23,28-29].

Macro-average metrics are calculated by averaging the metrics calculated for each label. Owing to the multiple and unbalanced classes involved in this task, macro-average metrics place more emphasis on the prediction of rare labels, which can better reflect the performance of the model in small classes. The micro average-recall rate and macro average recall rate are calculated as follows:

$$Micro - R = \frac{\sum_{l=1}^{|L|} TP_l)}{\sum_{l=1}^{|L|} TP_l + FN_l}, \quad (1)$$

$$Macro - R = \frac{1}{|L|}\sum_{l=1}^{|L|}\frac{TP_l}{TP_l + FN_l}, \quad (2)$$

where *TP* stands for true positive cases and *FN* stands for false negative cases. The calculation of accuracy is similar.

3.3. Summary

In summary, automated ICD coding is the process of assigning ICD codes y to health-related documents D. These documents commonly include EMRs, death certifications, clinical summaries, and medical reports.

More specifically, ICD coding requires the mapping of a series of diagnoses $Diag = [diag_1, diag_2, \ldots, diag_k]$ in document D to a series of codes $y = [icd_1, icd_2, \ldots, icd_k]$. Through representation learning, the document D is represented as a sequence of surrounding segments of diagnosis S_{diag_j}, where $1 \le j \le k$. S_{diag_j} can be narrowly understood as the disease name or a short description of the diagnosis result. In a broad sense, it is the set of all segments represented in the document that are related to $diag_j$. Take EMRs as an example: the direct diagnosis appears in the discharge diagnosis; however, the admission situation, admission diagnosis, discharge situation, examination results, and other segments may also be related to the diagnosis and can be used as surrounding information to assist coding.

Thus, the core process of this problem can be abstracted as the representation learning of documents to obtain diagnosis-related information and the classification of ICD codes from the diagnosis.

4. Developmental history of automated ICD coding

The development of automated ICD coding has taken place over decades. We divide it into three stages based on the techniques used. (1) In the rule-based stage, methods tried to extract and transfer the rules of coding specification to if-else logical programs to replace repetitive manual work [30]. After that, (2) traditional machine learning methods, such as support vector machine (SVM), were used to predict ICD codes from health-related documents [31–33]. At this stage, the focus was on constructing distinguishable features and finding suitable classifiers. With the development of neural networks, (3) deep learning methods emerged to tackle the ICD coding prediction, and the focus shifted to how to obtain better representations of health-related documents by designing different neural networks. In this section, we review some of these methods and summarize their performance on the publicly available Medical Information Mart for Intensive Care (MIMIC) dataset.

4.1. Stage 1: rule-based methods

Early automated ICD coding methods mainly relied on rules and expert experience [30,34] to mine and automatically generate coding rules by learning ICD guidelines. The ICD coding guidelines indicate the symptoms, signs, and corresponding constraints for each disease. Thus, many methods tried to convert the plain text into executable logical judgments so as to achieve automatic coding prediction. In addition, expanding the list of medical concepts in the coding guidelines was also an important way to enhance the effectiveness of rule-based coding methods. This could increase the concept coverage of health-related by using corresponding concepts in the guidelines with their abbreviations, synonyms, etc. According to the work of Farkas and Szarvas [30], the performance of rule-based methods on the Cincinnati Children's Hospital Medical Center (CCHMC) dataset could reach 90.26% (training set) and 88.93% (test set).

However, the disadvantages of the rule-based methods are their poor flexibility and portability. There are many common symptoms among different diseases; this can easily lead to over-coding and missing-code problems. More seriously, as the number of codes increases, it is easy for conflicts to arise among different rules. This problem can be manually screened and adjusted for in the CCHMC dataset as it only contains 45 ICD codes. However, if the number of codes rises to thousands or tens of thousands, rule-based methods may no longer be feasible.

4.2. Stage 2: traditional machine learning-based methods

With the emergence of machine learning, many studies began to use machine-learning-based methods to tackle the automated ICD coding task (Table 1). The key to traditional machine-learning-based methods is the construction and selection of features [35].

Suominen et al. [36] used feature engineering methods to extract features such as word segmentation results, medical concepts, and hypernyms, and then used a cascade of two classifiers: the regularized least squares (RLS) classifier and the RIPPER algorithm, to determine the final multi-label predictions. In experiments conducted on the CCHMC dataset, they achieved an 87.7% micro-averaged F1-score. Perotte et al. [32] took keywords obtained from Term Frequency-Inverse Document Frequency (TF-IDF) as features and proposed two models, FlatSVM and hierarchy-based SVM. The FlatSVM model treats each prediction independently. Marafino et al. [37] extracted n-gram features to build a binary SVM classifier for each disease. Elyne et al. [16] took both structured and unstructured clinical data into consideration, and a combined bag of words (BoW) from the unstructured text was selected as one of the features.

Most traditional machine learning methods train a separate classifier for each code. These methods can achieve good results when the number of codes to be predicted is small. For example, the CCHMC dataset used by Suominen et al. [36] contained 45 codes, and the experiment of Marafino et al. [37] was only conducted on four diseases, and positive and negative training samples were manually balanced. However, the number of ICD codes is rapidly increasing, and it is impractical to train thousands of classifiers in this way. Another drawback is that each disease is considered independently and the relationships between multiple ICD codes assigned to one EMR are ignored.

4.3. Stage 3: neural-network-based methods (Table 2)

With the development of deep learning, studies of automated ICD coding based on neural networks have gradually become mainstream.

CNN-based methods. The most classic examples are the CAML and DR-CAML models proposed by Mullenbach et al. [23]. These models use convolutional neural networks (CNNs) to automatically extract features in the discharge summaries, followed by a per-label attention mechanism. To obtain the probability of each label, sigmoid transformation is used for multi-label prediction. The models are simple but efficient and robust; they demonstrated breakthrough improvements on the MIMIC-II and MIMIC-III datasets. The micro-average F1 reached 0.457 (MIMIC-II), 0.633 (MIMIC-III-50), and 0.529 (MIMIC-III-Full).

Many subsequent studies were inspired by and improved on these models. Cao et al. [38] and Ji et al. [39] used a dilated CNN, which inserted "holes" into the filters, to extract non-continuous semantic information from the EMRs. Some researchers have argued that a flat and fixed-length convolutional architecture may not be able to learn a good document representation. Thus, multi-scale and variable-sized convolutional filters have been employed to capture various text patterns of different lengths [40–43].

These methods try to expand the receptive fields of filters and match different text patterns by changing the shape, structure, or scale of the filters, with the aims of obtaining a better text representation. Examples of these filters are shown in Figure 3. As well as multi-scale convolution, Li and Yu [42] and Ji et al. [39] added a residual block [44] on the convolution layers; this was also beneficial for the expansion of the receptive field.

Recurrent neural network (RNN)-based methods. In another series of studies, RNNs were employed to extract features from medical

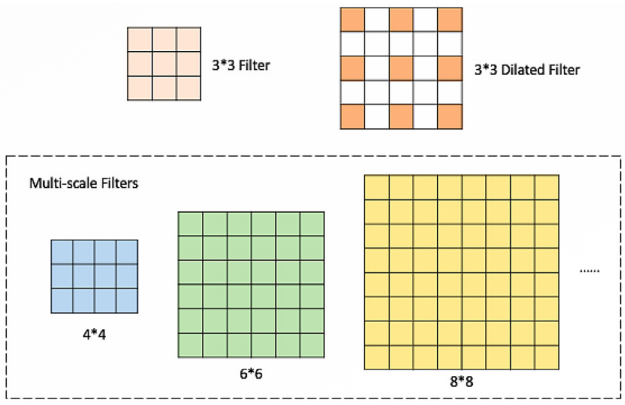

Figure 3. The different filters used in CNN-based methods.

texts. Shi et al. [45] used two-level long short-term memory (LSTM) networks (character-level and word-level) to obtain representations of the diagnosis description and ICD title, respectively. Baumel et al. [46] employed two layers of Gated Recurrent Unit (GRU) to encode the documents. The first layer encoded tokens to generate sentence representations, and the second layer encoded the sentences to obtain a whole-document representation. The attention mechanism was applied to both layers. Yu et al. [24] proposed a multi-layer attention bidirectional RNN (MA-BiRNN) model, in which one layer learned word-level features of a document, and the other layer learned word-level features.

It is challenging to model long sequences and realize parallel computing with RNNs; thus, the above works used multi-layer structures to decompose the documents and reduce the length of the sequence. Yu et al. [24] proposed, the average number of words per document was 610, and Shi et al. [45] only selected the diagnosis description from the discharge summary to encode. Another approach to decrease the length of the document is to extract the symptoms from the text and use LSTM to model the relation between symptom sequence and disease [47]. It can be inferred that a short text length is vital to the performance of RNN-based models.

Graph neural network (GNN)-based methods. GNNs have also been applied in automatic coding. The main method is to employ GNNs to learn the representation of ICD codes, as the codes are organized in a tree-based structure. Rios and Kavuluru [48] used the graph convolutional network (GCN) model to obtain features for each label, aiming to solve the few-shot and zero-shot prediction problems in ICD coding. Similarly, Du et al. [49] used the tree-based structure of ICD taxonomy to capture the dependency of codes. Cao et al. [25] also counted the co-occurrence frequency of ICD codes in EMRs to learn a code representation by GCN; this improved the incomplete prediction of coding. Moreover, Wang et al. [50] proposed two approaches to edge construction in the ICD code graph: point-wise mutual information between codes, and the TF-IDF value. Although they did not conduct experiments on the ICD coding task, their work provides a reference for graph construction. Gao et al. [51] proposed an unsupervised semantic-based heterogeneous graph representation method (SMP-Graph) that inductively enhanced the axis word-level, chapter-level, and document-level knowledge implicit in each piece of coding text.

A few studies have used knowledge graphs (KGs) to enhance the representation of medical texts or ICD codes. Teng et al. [52] built a KG with more than 1,500 nodes by fusing ICD-9 descriptions and medical-related data from the Freebase database. THe SDNE algorithm was used to encode the ICD codes, and the KG embedding helped to understand the terminology. By contrast, Chelladurai et al. [53] constructed an initial contextual graph with entities extracted from the clinical notes and those enriched by a pre-constructed external KG. Then, a GNN was employed to filter relevant nodes in the graph and contextualize the concepts, and an information retrieval system was used to obtain the final prediction. Wang et al. [54] constructed a heterogeneous graph from each EMR, enriched by the Wikipedia database. They pre-trained a graph encoder with two graph contrastive learning schemes.

Pre-trained language model (PLM)-based methods. In recent years, PLMs have achieved state-of-the-art results on multiple NLP natural language processing tasks. A few studies have employed PLMs to enhance the ability of text representation. Owing to the occurrence of specialized terms and expressions, BERT [55] is not well-suited to medical texts; however, the pre-training models specifically trained on biomedical texts, such as BioBERT [56], ClinicalBERT [57], and PubMedBERT [58], may better help to understand such texts.

Zhang et al. [59] trained BERT on EMRs and used it to encode text. They extended the maximum sequence length to 1,024; 90% of documents in their dataset did not contain more than 800 tokens, and the documents that exceeded 1,024 tokens were split into segments. The highest predicted probability per ICD code across segments was used as the note-level prediction. As shown in Table 4, the average token number per document in common EMR datasets is no less than 1,100, and the maximum token number of the MIMIC dataset exceeds 2,500. Therefore, splitting input text into segments is inevitable. Pascual et al. [60] directly employed PubMedBERT [47] as an encoder and explored different ways to split text. A decoder was also needed to combine the representations of each chunk. Surprisingly, the experiments illustrated that splitting the text into chunks of the same size resulted in the worst performance; even taking only the first 512 tokens or last 512 tokens was a better approach.

Based on the results obtained with the above PLM-based methods, although PLMs show outstanding ability in understanding medical language, they do not perform well on ICD coding tasks. This may account for few PLM-based methods being developed for this task. The poor performance may be affected by the long length of input medical documents. PLMs can often only process 512 tokens at once, which is less than the token number of most documents used in this task, and the necessary splitting of text has a negative impact on the final performance.

4.4. Summary

In summary, the automated ICD coding task has been a hot research topic in recent years, and many excellent methods and models have been proposed. In addition, many new sub-tasks and processing methods have been developed to assist automated ICD coding tasks, such as the prediction of the number of labels [15].

To obtain a more intuitive understanding of the progress that has been made, we summarize the results of state-of-the-art methods on the MIMIC datasets in Table 3, as MIMIC is the most widely used public dataset. To be specific, there are three datasets: MIMIC-II, MIMIC-III-50, and MIMIC-III-Full.

In the first block of the table, the performances of three typical models are reported, as implemented by Mullenbach et al. [23]. These text classification methods did not show satisfactory performance in general, and there was a significant gap in the F1 scores of these typical models and those obtained with improved deep learning models. The best Micro-F1 scores on MIMIC-II, MIMIC-III-50, MIMIC-III-Full were 0.498, 0.725, and 0.575, respectively. The Fusion method proposed by Luo et al. [61], ISD proposed by Zhou et al. [62], LAAT&JointLAAT proposed by Vu et al. [63], and MSATT-KG proposed by Xie et al. [64] showed the most competitive performance.

It can also be inferred from the results shown in Table 3 that the medical field is knowledge-intensive, and it is more difficult to capture features from medical text than general texts, hence the need to design more elaborate network structures or use more external knowledge. Most of the methods listed in Table 3 have been introduced in stages 2 and 3 above; the others will be described in detail in Section 6.

Table 1 Traditional machine-learning-based methods

researches	Models	Methods/Features
(Larkey&Croft, 1996) [5]	K-nearest-neighbor; Bayes classifier;relevance feedback	Terms and phrases
(Suominen et al., 2007) [36]	RLS and RIPPER	Word segmentation,medical concepts, and hypernyms
(Perotte et al., 2013) [32]	FlatSVM and hierarchy-based SVM	TF-IDF
(Marafino et al., 2014) [37]	SVM	N-gram
(Elyne et al., 2016) [16]	Naive Bayes classifier and random forests	BoW;TF-IDF;demographic data;laboratory results, etc.

Table 2 Neural network-based methods

Approaches	Model descriptions		Input medical text (all text in the document included?)	Output layer
	CNN-based			
CAML & DR-CAML [23]	CNN + label-wise attention		√	Sigmoid activation
DCANM [38]	Dilated CNN + N-gram matching mechanism		√	Sigmoid activation
MVC-LDA & MVC-RLDA [40]	Multi-ScaleCNN	Max pooling across the channels	√	Sigmoid activation
Ontological attention [41]		Ontological attention and mapping	√	Sigmoid activation
MultiResCNN [42]		Residual convolutional layer	√	Sigmoid activation
EnCAML [43]		Concatenate feature maps horizontally	√	Sigmoid activation(variable threshold)
	RNN-based			
C-LSTM-ATT [45]	LSTM	English letter-level + word-level	Diagnosis descriptions	Sigmoid activation
MA-BiRNN [24]		Chinese character-level + word-level	Feature words	Sigmoid activation
HA-GRU [46]	GRU	Token-level + sentence-level	√ (Split into sentences)	Sigmoid activation
	GNN-based			
ZAGCNN [48]	GCNs is employed to learn code representation	Use both label descriptors and structure	√	Sigmoid activation
HyperCore [25]		Graph constructed by codes co-occurrence	√	Sigmoid acivation
MSResGCN [49]		Label graph: ICD tree-based structure	√	Sigmoid activation
GrabQC [53]	GNN is to encode documents	Recognize named entities and link to KG	√	Information retrieval system
	PLMs-based			
BERT-XML [59]	Train BERT model on EMRs; most documents do not need to be split as the maximum sequence length is extended to 1,024		√	Sigmoid activation
BERT-ICD [60]	PubMedBERT; five text-splitting strategies		√	Not mentioned

5. Commonly used datasets for automated ICD coding

Many datasets in various languages have been exploited and made public. We list 17 commonly used datasets in Table 4 and report their detailed statistical information. Similar to most previous studies, we only performed statistical analyses on the unstructured text information in the MIMIC dataset.

The most commonly used dataset is MIMIC [66–67]. It collects comprehensive clinical data from tens of thousands of intensive care unit (ICU) patients from 2001. The data include discharge summaries, radiology reports, laboratory measurements, microbiology cultures, medication prescriptions, vital signs, and other numerical and textual data. The diagnoses in the MIMIC dataset are all labeled by ICD-9 codes. MIMIC-III-50 is a subset of MIMIC-III-Full, including the EMRs labeled by at least one of the top 50 most frequently used codes. MIMIC-IV [67] is the latest version of the MIMIC datasets. Unlike MIMIC-III, MIMIC-IV is grouped into several modules, i.e., Core, Hosp, ICU, ED, CXR, and Note. The note module, which contains free-text clinical notes, has not been made publicly available.

The CCHMC dataset [68], also called the Computational Medicine Center dataset, was widely used in the early period of research on this topic and collects radiology reports from the CCHMC. It exploits the “majority” rule, and each document is labeled by only one code.

The remainder of the datasets are provided by various hospitals and medical institutions. UKLarge and UKSmall [15] collected EMRs from the University of Kentucky Medical Center between 2011 and 2012. The Xiangya dataset [24] is derived from three affiliated hospitals of Central South University. The CN-Full dataset [38] is derived from a cooperative medical institution. The EMRs of UZA [16] are from the Antwerp University Hospital.

Conference and Labs of the Evaluation Forum (CLEF) has evaluated ICD coding of health-related documents every year from 2017 to 2020, and its datasets are mainly based on French, Spanish, German, and other European languages. The CDC (Center for Disease Control) dataset and CepiDc-2017 dataset [69] are the specific datasets used in CLEF eHealth 2017. The CepiDc-2018, CLEF-Italian, and CLEF-Hungarian datasets [26] were publicly available for use in CLEF eHealth 2018. They all collect electronic death certificates as the documents to be assigned ICD codes.

THe CLEF-German dataset [27], released by CLEF in 2019, consists of NTS of animal experiments. Each NTS contains a title, uses (goals) of the experiments, possible harms caused to the animals, and comments about replacement, reduction, and refinement [70]. It is annotated with chapters or group codes from the ICD-10 German Modification 2016 version. For example, the entorhinal incoming of epilepsy the Entorhinal Afferents in Epilepsy experiment carried out on mice is labeled with VI

Table 3 The results reported on the MIMIC-II, MIMIC-III-Full, and MIMIC-III-50 datasets

Models	MIMIC-II					MIMIC-III-50					MIMIC-III-Full					
	AUC		F1		*P*@8	AUC		F1		*P*@5	AUC		F1		*P*@8	*P*@15
	Mac	Mic	Mac	Mic		Mac	Mic	Mac	Mic		Mac	Mic	Mac	Mic		
Logistic regression [23]	0.690	0.934	0.025	0.314	0.425	0.829	0.864	0.477	0.533	0.546	0.561	0.937	0.011	0.272	0.542	0.411
CNN [23]	0.742	0.941	0.030	0.332	0.388	0.876	0.907	0.576	0.625	0.620	0.806	0.969	0.042	0.419	0.581	0.443
Bi-GRU [23]	0.780	0.954	0.024	0.359	0.420	0.828	0.868	0.484	0.549	0.591	0.822	0.971	0.038	0.417	0.585	0.445
Flat SVM [32]	-	-	-	0.211	-	-	-	-	-	-	-	-	-	-	-	-
Hierarchy-based SVM [32]	-	-	-	0.293	-	-	-	-	-	-	-	-	-	-	-	-
C-LSTM-ATT [45]	-	-	-	-	-	-	0.900	-	0.532	-	-	-	-	-	-	-
C-MemNN [28]	-	-	-	-	-	0.833	-	-	-	0.420	-	-	-	-	-	-
HA-GRU [46]	-	-	-	0.366	-	-	-	-	0.366	-	-	-	-	-	-	-
CAML [23]	0.820	0.966	0.048	0.442	0.523	0.875	0.909	0.532	0.614	0.609	0.895	0.986	0.088	0.539	0.709	0.561
DR-CAML [23]	0.826	0.966	0.049	0.457	0.515	0.884	0.916	0.576	0.633	0.618	0.897	0.985	0.086	0.529	0.690	0.548
LEAM [20]	-	-	-	-	-	0.881	0.912	0.540	0.619	0.612	-	-	-	-	-	-
MA-BiRNN [24]	-	-	-	-	-	-	-	-	-	-	-	-	-	0.420	-	-
MSATT-KG [64]	-	-	-	-	-	0.914	0.936	0.638	0.684	0.644	0.910	**0.992**	0.090	0.553	0.728	0.581
KAICD [65]	-	-	-	-	-	-	-	-	-	-	-	-	-	0.462	-	-
HyperCore [25]	0.885	0.971	0.070	0.477	0.537	0.895	0.929	0.609	0.663	0.632	0.930	**0.989**	0.090	0.551	0.722	0.579
DACNM [38]	-	-	-	-	-	0.890	0.916	0.579	0.641	0.616	-	-	-	-	-	-
DCAN [39]	-	-	-	-	-	0.902	0.931	0.615	0.671	0.642	-	-	-	-	-	-
MSResGCN [49]	-	-	-	-	-	-	-	-	-	-	0.877	0.983	0.076	0.539	0.722	0.568
MultiResCNN [42]	0.850	0.968	0.052	0.464	0.544	0.899	0.928	0.606	0.670	0.641	0.910	0.986	0.085	0.552	0.734	0.584
LAAT [63]	0.868	**0.973**	0.059	0.486	0.550	0.925	**0.946**	0.666	0.715	0.675	0.919	**0.988**	0.099	**0.575**	0.738	**0.591**
JointLAAT [63]	0.871	**0.972**	0.068	0.491	0.551	0.925	**0.946**	0.661	0.716	0.671	0.921	**0.988**	0.107	**0.575**	0.735	**0.590**
BERT-ICD [60]	-	-	-	-	-	0.845	0.887	-	-	-	-	-	-	-	-	-
G_coder [52]	-	-	-	-	-	-	0.933	-	0.692	0.653	-	-	-	-	-	-
ISD [62]	**0.901**	**0.977**	**0.101**	**0.498**	**0.564**	**0.935**	**0.949**	**0.679**	0.717	**0.682**	**0.938**	**0.990**	**0.119**	0.559	**0.745**	-
Fusion [61]	-	-	-	-	-	**0.931**	**0.950**	**0.683**	**0.725**	**0.679**	0.915	**0.987**	0.083	0.554	0.736	-

The best results are underlined, and those that are less than 0.05 from the best result are marked in bold. This table only compares methods that use the same text pre-processing technique; thus, the results of experiments conducted on the MIMIC dataset using different pre-processing techniques are not listed.

Table 4 Datasets commonly used for ICD coding

Language	Dataset	Public	Data Format	Document Type	ICD Version	# Documents	Avg Token # per Document	Avg Labels# per Document	Total # Labels
English	MIMIC-II	√	①②	EMRs	ICD-9	20,533	1,138	9.2	5,031
	MIMIC-III-Full	√	①②	EMRs	ICD-9	47,724	1,485	15.9	8,922
	MIMIC-III-50	√	①②	EMRs	ICD-9	8,067	1,530	5.7	50
	CCHMC	√	②	Radiology reports	ICD-9-CM	1,954	21	1	45
	CDC	√	②	Death certificate	ICD-10	13,330 (train) 6,665 (test)	6.8 6.4	3.0 2.8	1,256 900
	UKLarge	✗	②	EMRs	ICD-9-CM	71,463	5,303	-	1,231
	UKSmall	✗	②	EMRs	ICD-9-CM	1,000	2,088	-	56
Chinese	Xiangya	✗	②	EMRs	ICD-10	7,732	610	3.6	1,177
	CN-Full	✗	②	EMRs	ICD-10	50,678	621	4.3	6,200
	CN-50	✗	②	EMRs	ICD-10	36,758	655	2.6	50
Dutch	UZA	✗	①②	EMRs	ICD-9-CM	56,641	-	-	23,727
French	CepiDc-2017	√	②	Death certificates	ICD-10	65,844 (train) 27,850 (test)	17.9 17.8	4.1 4.0	3,233 2,363
	CepiDc-2018	√	②	Death certificates	ICD-10	125,384 (train) 11,932 (test)	-	-	-
Italian	CLEF-Italian	√	②	Death certificates	ICD-10	14,502 (train) 3,618 (test)	-	-	-
Hungarian	CLEF-Hungarian	√	②	Death certificates	ICD-10	84,703 (train) 21,176 (test)	-	-	-
German	CLEF-German	√	②	NTS of animal experiments	ICD-10 German modification 2016 version	8,386 (train & dev) 407 (test)	369.17	2.6	233
Spanish	CLEF-Spanish	√	②	EMRs	Spanish version of ICD10-CM and ICD10-PCS	1,000	396.99	18.4	3,427

①:Structured and ②Unstructured.

and G40-G47, where VI is the code of the Nervous System Disease the Diseases of the Nervous System chapter and G40-G47 is the code of the Transient and Sudden Illness Episodic and Paroxysmal Diseases of the Nervous System section.

The CLEF-Spanish dataset [71], also called the CodiEsp dataset, is the specific dataset used in CLEF eHealth 2020: 1,000 EMRs are used for the training set, development set, and test set. In addition, 2,751 unlabeled EMRs are provided as a background set. The EMRs include documents labeled by both diagnosis codes and procedure codes. Ignoring the type of document, the average number of labels for 1,000 EMRs is 18.4, as reported in Table 4. However, the average number of labels for each procedure document is 8.2, whereas the number for diagnosis documents is 25.27.

The majority of the datasets are in English, followed by Chinese. The rest are in various European languages.

Most of the English-language datasets are publicly available and used widely. Among them, the MIMIC dataset is the most frequently used dataset in published studies. The European-language-based datasets are public, whereas the Chinese-language datasets are all non-public.

The datasets consist of health-related documents, mainly EMRs, radiology reports, death certificates, and NTS reports. The EMRs are mainly collected from hospitalized patients. The document lengths of the death certificates and radio-graphic reports are relatively short, with an average length of between 6.4 and 21 English words. The average length of English EMRs ranges from 1,138 to 5,303 words, and the average length of Chinese EMRs is about 600 words.

Regarding the total number of ICD codes in the datasets, with the exception of the CCHMC dataset, which only contains 45 ICD codes, most of the datasets have no fewer than 1,000 ICD codes. The MIMIC-III dataset contains the most codes, with a total number of 8,922. For datasets with large numbers of tags, such as the MIMIC-III-Full dataset, CN-Full dataset, and UKLarge dataset, some codes with a higher frequency are selected to generate a corresponding subset: the MIMIC-III-50, CN-50, and UKSmall datasets.

6. Challenges and existing solutions in automated ICD coding

Automated ICD coding is an important fundamental task in the domain of smart medicine. Although many studies have attempted to address this task using many different approaches, varying from rule-based to neural-network-based methods, and achieved great improvements in performance on several evaluation metrics, there are still many challenges to be solved. These include the high dimension of the label space, severe imbalances in numbers of samples for each disease/code, irregular language expressions, and inevitable noise in EMRs. These problems become more prominent in the practical clinical environment, and there is a long way to go before a robust, reliable, and explainable AI infrastructure is constructed. Thus, we summarize four challenges of automated ICD coding tasks and and also review some outstanding work that attempts to break through the bottleneck.

6.1. Large label space

The large number of ICD codes leads to a huge label space for the coding prediction task. ICD-9-CM includes about 13,500 diagnostic codes and 4,000 procedure codes, each of which contains a maximum of four digits. In the next version, ICD-10-CM, the number of diagnostic codes increases to more than 70,000, the number of procedure codes increases to 72,000, and the number of digits per the code reaches seven [3]. The large label space leads to difficulties with coding prediction, and it is expected to become larger still with further iterations of the ICD.

Most approaches to deal with this problem use the three characteristics summarized in Section 2, that is, they make use of the hierarchical architecture of the ICD taxonomy, explore the relationships between codes, and conduct reasonable statistical analysis.

Given the characteristic of inheritance of ICD codes, Vu et al. [59] proposed the JointLAAT model to learn the hierarchical relationship between codes through a hierarchical joint learning mechanism. Falis et al. [40] proposed ontological attention to capturing semantic concepts in ICD code prediction from the clinical text by considering the hierarchical structure of the coding; three-layer coding is used as a label-wise approach to improve performance.

Regarding the mutual exclusion of ICD codes, various methods exist for preventing sibling nodes from being selected simultaneously. Subotin and Davis [72] considered simultaneous prediction as an asynchronous process and introduced conditional probability to learn the probability of code B being assigned given that code A has been assigned to the medical record. Pengtao et al. [73] used the sequence tree-based LSTM network to capture the relationship between codes, effectively preventing a common choice of sibling codes. Moreover, Cao et al. [25] converted the ICD taxonomy from Euclidean space to hyperbolic space, increasing the distance between nodes at the same level in hyperbolic space to reflect the mutual exclusion relationship of sibling nodes.

The co-occurrence of coding has also been noted in recent years. Cao et al. [25] calculated the frequency of co-occurrence of codes in EMRs and built a code graph where the edge weight represented the frequency. Graph convolutional networks have been utilized to learn the co-occurrence relationship between codes and obtain a representation of the code. In addition to considering the co-occurrence relationship during modeling, it is also useful to consider it in the post-processing stage to improving the final result. Tsai et al. [74] regarded the automated coding task as a two-stage task, where the first stage is model prediction, and the second stage is re-ordering the predicted codes based on the label distribution of the dataset. The whole process is equivalent to adding post-processing to the original model. Experiments show that adding constraints of the co-occurrence relationship to three existing models (CAML, MultiResCNN, and LAAT) can improve the effect of the original model.

6.2. Unbalanced label distribution

The extremely unbalanced distribution of codes is the second biggest challenge in the task of automated ICD coding. A few codes appear very frequently in EMRs, such as respiratory infections and coughs, whereas most of the codes have a very low frequency of appearance; this leads to the serious long-tails phenomenon [6]. Some studies also refer to it as the Zipffe distribution of the ICD codes in EMRs, and the datasets with long tails are called “power-law datasets” [75].

We investigated some datasets to explore the specific distribution of ICD codes; the results are shown in Figure 4. In the MIMIC-III dataset, 10% of ICD codes appear in 85% of the data, whereas 22% of codes appear fewer than two times, and about 5,000 labels appear between one and ten times. More seriously, more than 50% of diagnostic and procedure codes, around 17,000, never appear in the dataset [45,60].

The situation with respect to the Chinese datasets is even more serious. In CN-Full and CN-50 [38], 32.8% of codes only appear once, 46.9% of codes appear less than two times, and 74.6% of codes appear between one and ten times. The most high-frequency code (Essential hypertension) appears 9,544 times in EMRs.

The highly frequent occurrence of a minority of codes is consistent with the situation in real medical settings. Most patients in a hospital suffer from common diseases; a few additional diseases may be encountered occasionally, and many rare diseases in the ICD taxonomy may never appear. In addition, some ICD codes recently introduced in the latest version may be rarely or never used.

This problem is likely to be missed by data-driven machine learning methods because the wrong prediction of infrequent labels has little effect on the final evaluation metrics. However, it is important for

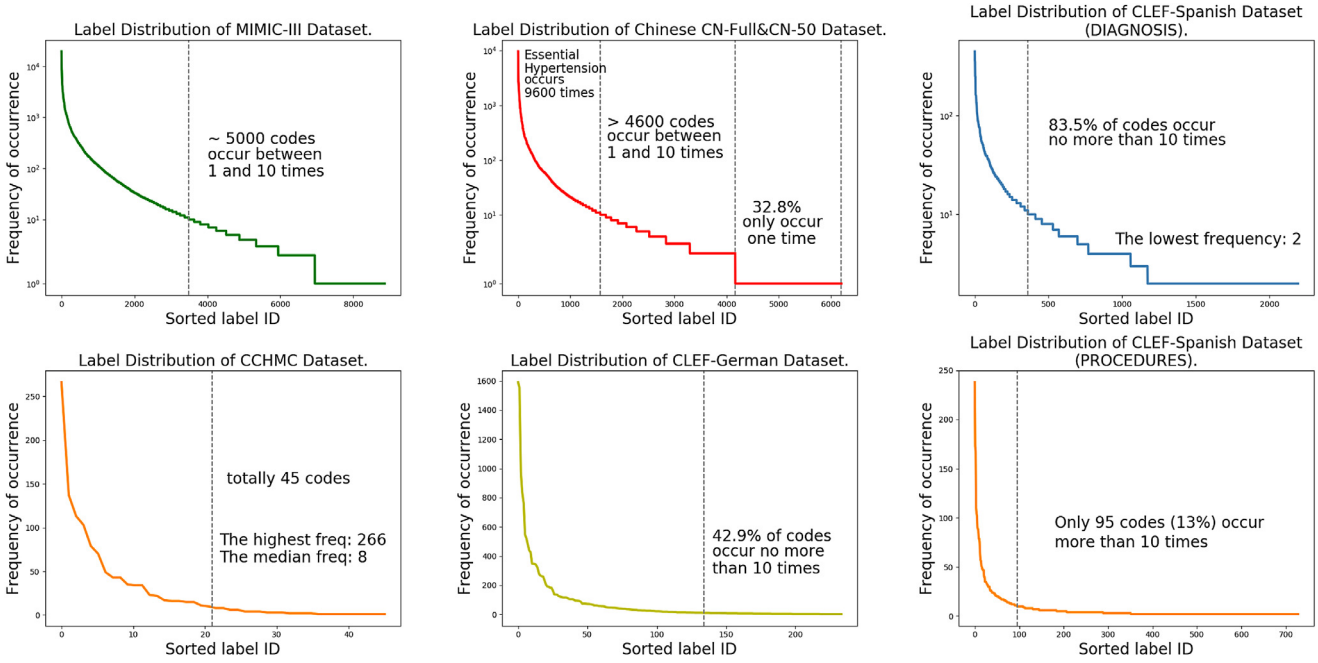

Figure 4. The label frequency distribution in different datasets.

both medical records management and medical reimbursement for such rare labels to be correctly predicted. The main deep-learning-based methods intended to deal with this challenge can be summarized as follows.

Sampling. The general strategy to address the problem of unbalanced label distribution in machine learning is to improve it through methods such as random under- or over-sampling, synthetic training sample generation, and cost-sensitive learning. Kavuluru et al. [15] constructed an "optimal training set" for each label by under-sampling negative samples. However, this method converts the multi-label classification problem into multiple binary classification problems, which do not have universal adaptability.

Introducing the name and description of the disease. The disease name is the most intuitive piece of information. Shi et al. [43] encoded the disease name (long title) encoded by the ICD to generate a label representation. Wu et al. [58] used two different networks to extract the features of disease names and the medical record text. The medical record text is extracted using a multi-scale CNN, and the disease name is extracted using Bi-GRU. For codes whose disease names themselves do not provide enough information, a deeper understanding can be obtained from the disease descriptions given in the ICD coding system. Rios and Kavuluru [45] used a GCN model to improve the few-shot and zero-shot prediction problems in ICD coding, adding the description information of the code and using graph convolution to learn according to the hierarchical structure of the code itself, instead of the random initialization vector in the CAML model [23]. Moreover, Pengtao et al. [73] used the coding description and also introduced adversarial learning to solve the problem of inconsistent styles between the disease description and EMR text.

Introducing external knowledge. To the best of our knowledge, the earliest work using external knowledge is that of [28], who stored the first paragraph of the text and corresponding "Signs and symptoms" section in Wikipedia for each diagnosis, and employed condensed memory networks to enable interaction of clinical notes with the knowledge base. Wikipedia was also used as an external knowledge source in the work of [76]. In addition to external data sources, the hierarchical ICD taxonomy itself can also be used as knowledge, as the high-level codes can provide general information for its lower-frequency low-level codes. Tsai et al. [77] used the hierarchical category knowledge of the diagnostic code to allow different labels in the same high-level category to share semantics in order to solve the problem of label imbalance. They took the high-level coding prediction as the auxiliary task to train together with the low-level coding task. Zhou et al. [60] proposed a shared attention representation, extracting the shared features of low-frequency coding and high-frequency coding, to overcome the difficulty of learning accurate representations of rare codes owing to insufficient data. Wang et al. [78] used both the "Signs and symptoms" section in Wikipedia and the ICD hierarchical structure.

Moreover, transfer learning can be used to transfer domain knowledge. Zeng et al. [79] utilized automatic MeSH indexing as an extra auxiliary task to learn domain knowledge and transferred it to an automatic ICD coding task. MeSH is a medical literature data source that contains tens of millions of samples.

6.3. Long text of documents

Generally speaking, health-related documents may be very long; this applies especially to EMRs, which include past history, current history, examination results, diagnosis, etc., and use many complicated medical expressions and professional terms. Moreover, many abbreviations, aliases, and non-standard terms, as well as some misspelled words, may appear in medical records; thus, EMRs represent high-noise and high-sparsity text. Therefore, modeling of long medical texts will inevitably be affected by redundancy or errors in information, potentially leading to important information being missed.

To our knowledge, few approach have been proposed to specifically solve this problem. Some specific mechanisms may be invoked to improve the representation of long texts. Zhou et al. [60] used a self-distillation learning mechanism to deal with the noise problem in long texts. Their teacher model used the description of the target code, whereas the student model used the original text with noise to learn the ability to extract key information from a long text. Luo et al. [61] compressed the features obtained with filters by attention; the sliding windows of filters produce redundant information, so pooling is required to distinguish the important adjacent phrases and filter out noise. These attempts indicate some possible directions for that could lead to breakthroughs regarding this challenge, but there is a long way to go before satisfactory modeling of long documents is achieved.

6.4. Interpretability of coding

One of the drawbacks of deep learning models is their poor interpretability. However, the interpretability of the automated ICD coding task as a decision-making assistance task is essential for its clinical applications. In other words, it is necessary to provide corresponding supporting information and decision-making assistance when predicting ICD codes to improve the interpretability of the model [29].

The attention mechanism is the most intuitive and widely used method to provide interpretability, as it can infer which words or fragments the model pays more attention to when predicting. Mullenbach et al. [23] used a per-label attention mechanism to learn the importance weight of words in a document; the greater the weight, the more relevant the word to the current label. Baumel et al. [44] used sentence-level attention to identify sentences related to each label in a document; sentences represent more coarse-grained and more complete information than words.

In addition to the attention mechanism, Cao et al. [38] proposed the DCANM model, which uses an n-gram matching mechanism to obtain continuous word semantics and uses dilated convolution to obtain non-continuous word semantics. The matched words are seen as interpretable information. More intuitively, Duque et al. [80] used the TF-IDF technique and medical tagger tools to extract relevant words, phrases, and medical concepts for each code from EMRs to their constructed knowledge base. The testing of the model included a mapping and ranking process to complete the assignment.

7. Applications of automated ICD coding

7.1. Assisting professionals in decision-making

The most straightforward application of automated ICD coding tasks is to assist professional coders to finish the assignment decision-making process. This frees humans from repetitive and time-consuming work and improves the efficiency and quality of coding. Moreover, the coding system can also be used in the auto-checking and error correction of already-coded health-related documents to control the quality of ICD code assignment.

7.2. Statistics and analysis on disease and death

Analysis of health information statistics has an irreplaceable role in the development of the world's medical and healthcare systems. The introduction of ICD codes in the 19th century was intended to classify and count the causes of deaths from disease. Incidence, mortality, and survival are still key indicators for hospitals, national medical departments, the World Health Organization, and other institutions.

Through ICD codes, statisticians can clearly analyze the occurrence and development of diseases in a certain hospital, a certain community, and a certain area; the types of diseases in a certain hospital, the types of diseases, and the types of outpatient presentations; causes of death; and reasons for injuries and poisonings. Further statistical analyses can be implemented on outpatient data, hospitalization data, medical data, management data, medicinal materials data, and economic data, and the results can be used to guide decision-making.

7.3. Medical data standardization and sharing

Digital and intelligent medical treatment has become an inevitable trend in the development of medical management. ICD coding provides a standard medical information system, which is a prerequisite for digital medical care. The first problem that needs to be solved is the coding problem of medical records, so as to realize the standardization of diagnostic and treatment information and further achieve data sharing. The widespread use of ICD coding enhances communication and sharing between hospitals, regions, and countries.

7.4. Medical expenses payment

Control of unnecessary medical expenses payments is an important application of ICD coding. DRGs based on ICD coding represent a vital means of achieving effective management of medical quality and costs. DRGs are used to group cases with similar clinical processes and similar cost consumption based on comprehensive consideration of the patient's age, gender, length of stay in the hospital, clinical diagnosis, illness, surgery, comorbidities, complications, etc. A system that is divided into the same group for management.

The implementation of DRGs can reduce the loss of medical payments, ensure that the main diagnostic and treatment measures paid for by medical insurance are focused on patients who need them most, and guide the use of limited medical funds to protect more people.

7.5. Medical record management

In EMRs, the first page of the inpatient medical record is a concise description of the patient's diagnosis and treatment. Medical record management is an important part of hospital management as it provides an abundant resource for clinical decisions, teaching, and scientific research.

The application of an ICD automatic coding system places higher requirements on doctors' medical record-writing; the clarity and accuracy of the main diagnosis on the first page of the medical record are particularly important. Moreover, the sharing of coding information between clinical departments and medical record systems can help doctors to fill in the information on the first pages of medical records, ensure the quality of imported information, and promote improvements in the quality of hospital medical record management.

7.6. Hospital evaluation management

Health-related document databases can be searched by disease code for medical record content including disease diagnosis, pathological diagnosis, surgical and operation incision classification, hospital infection diagnosis, death of patients, etc. This information can be retrieved for a specific time period to evaluate the hospital's performance during that time. Hospital management personnel and medical quality management personnel can obtain real global data on hospital management through this path, which is beneficial to the management and standardization of their practice.

Moreover, with the continuous advancement of DRGs, indicators such as the number of groups of DRGs, case-mix index, time consumption index, cost consumption index, low-risk group mortality, and non-entry status are gradually being utilized to evaluate the performance of the hospital, and it makes it possible to compare the performance between different hospitals and different departments.

8. Discussion and future prospects

In this paper, we have introduced the importance and necessity of the ICD coding task, reviewed the research progress with respect to this task in recent years, and summarized the work on how to translate this problem into a learnable task. The development history of automated ICD coding can be divided into three stages. The two key points when solving this problem are the representation learning of documents to obtain diagnosis-related information and the classification of ICD codes from the diagnosis. We have also summarized the main challenges to be overcome.

For future studies, making this task a knowledge-driven task is a promising direction. As domain knowledge has been accumulated over many years, such as in medical KGs, using attributes of this knowledge to enhance our understanding of disease connotations may be beneficial to coding prediction.

Another problem to be considered is practicality of the clinical applications in medical systems. Potential solutions include building coding models of different scales or granularities to meet the actual coding needs of different scenarios; implementing conversions between different ICD versions (different languages, different hospitals), etc; and helping the iteration of new and old versions.

Conflicts of interest statement

The authors declare that there are no conflicts of interest.

Funding

This research was supported by Beijing Municipal Natural Science Foundation (Grant No. M22012) and BUPT Excellent Ph.D. Students Foundation (Grant No. CX2021122).

Author contributions

Chenwei Yan: Conceptualization, Writing – original draft, Writing – review & editing, Data curation, Formal analysis. **Xiangling Fu:** Conceptualization, Writing – review & editing. **Xien Liu:** Conceptualization, Writing – review & editing. **Yuanqiu Zhang:** Data curation, Formal analysis. **Ji Wu:** Conceptualization, Writing – review & editing. **Qiang Li:** Writing – review & editing.

References

[1] Manchikanti L. Implications of fraud and abuse in interventional pain management. Am Soc Intervent Pain Phys 2002;5(3):320–37.
[2] Dee L. Consultant report-natural language processing in the health care industry. 2007.
[3] Kaur R. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy: Association for Computational Linguistics; 2019. p. 1–9. doi:10.18653/v1/P19-2001.
[4] Yang Y, Chute CG. Proceedings of the Annual Symposium on Computer Application in Medical Care. California: IEEE Computer Society; 1994. p. 157–61.
[5] Larkey LS, Croft WB. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '96. New York, NY, USA: Association for Computing Machinery; 1996. p. 289–97. doi:10.1145/243199.243276.
[6] Zhang D, He D, Zhao S, et al. BioNLP 2017. Vancouver, Canada,: Association for Computational Linguistics; 2017. p. 263–71. doi:10.18653/v1/W17-2333.
[7] Aden. Medical record disease classification and coding defect analysis report in 2019, 2019,
[8] Kumar V, Recupero DR, Riboni D, et al. Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. IEEE Access 2020;9:7107–26. doi:10.1109/ACCESS.2020.3043221.
[9] Kumar V, Mishra BK, Mazzara M, et al. Prediction of malignant & benign breast cancer: a data mining approach in healthcare applications. arxiv 2019. doi:10.48550/arXiv.1902.03825.
[10] Dessì D, Helaoui R, Kumar V, et al. TF-IDF Vs word embeddings for morbidity identification in clinical notes: an initial study. arxiv 2021. doi:10.5281/zenodo.4777594.
[11] Abhyankar S, Demner-Fushman D, Callaghan FM, et al. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inf Assoc 2014;21(5):801–7. doi:10.1136/amiajnl-2013-001915.
[12] Friedman C, Shagina L, Lussier Y, et al. Automated encoding of clinical documents based on natural language processing. J Am Med Inf Assoc 2004;11(5):392–402. doi:10.1197/jamia.M1552.
[13] Subotin M, Davis A. Proceedings of BioNLP 2014. Baltimore, Maryland: Association for Computational Linguistics; 2014. p. 59–67. doi:10.3115/v1/W14-3409.
[14] Rizzo SG, Montesi D, Fabbri A, et al. In: Ambite J-L, editor Data Integration in the Life Sciences. Cham: Springer International Publishing; 2015. p. 147–61.
[15] Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artif Intell Med 2015;65(2):155–66. doi:10.1016/j.artmed.2015.04.007.
[16] Scheurwegs E, Luyckx K, Luyten L, et al. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inf Assoc 2016;23:e11–19. doi:10.1093/jamia/ocv115.
[17] Chen Y, Lu H, Li L. Automatic icd-10 coding algorithm using an improved longest common subsequence based on semantic similarity. PLoS ONE 2017;12(3):e0173410. doi:10.1371/journal.pone.0173410.
[18] Mario A, Raquel M, Victor F, et al. ICD-10 Coding based on semantic distance: lsi uned at clef ehealth 2020 task 1. Proc Conf Labs Evaluat Forum 2020;2696.
[19] Ning W, Yu M, Zhang R. A hierarchical method to automatically encode chinese diagnoses through semantic similarity estimation. BMC Med Inform Decis Mak 2016;16:30. doi:10.1186/s12911-016-0269-4.
[20] Wang G, Li C, Wang W, et al. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 2321–31.
[21] Atutxa A, de Ilarraza AD, Gojenola K, et al. Interpretable deep learning to map diagnostic texts to icd-10 codes. Int J Med Inform 2019;129:49–59. doi:10.1016/j.ijmedinf.2019.05.015.
[22] Xu K, Lam M, Pang J, et al. In: Kale DC, Ranganath R, Wallace BC, editors Proceedings of the Machine Learning for Healthcare Conference, MLHC 2019, Ann Arbor, Michigan, USA. PMLR; 2019. p. 197–215.
[23] Mullenbach J, Wiegreffe S, Duke J, et al. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018. p. 1101–11.
[24] Yu Y, Li M, Liu L, et al. Automatic ICD code assignment of chinese clinical notes based on multilayer attention birnn. J Biomed Inform 2019;91:103114. doi:10.1016/j.jbi.2019.103114.
[25] Cao P, Chen Y, Liu K, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 3105–14.
[26] Suominen H, Kelly L, Goeuriot L, et al. In: Murtagh F, Nie JY, Soulier L, editors Experimental IR Meets Multilinguality, Multimodality, and Interaction. Cham: Springer International Publishing; 2018. p. 286–301.
[27] Kelly L, Suominen H, Goeuriot L, et al. In: Rauber A, Müller H, Losada DE, editors Experimental IR Meets Multilinguality, Multimodality, and Interaction. Cham: Springer International Publishing; 2019. p. 322–39.
[28] Prakash A, Zhao S, Hasan S, et al. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; 2017. p. 3274–80.
[29] Vani A, Jernite Y, Sontag D. Grounded recurrent neural networks. arXiv preprint arXiv:170508557 2017.
[30] Farkas R, Szarvas G. Automatic construction of rule-based icd-9-cm coding systems. BMC Bioinformatic 2008;9 Suppl 3(Suppl 3):S10. doi:10.1186/1471-2105-9-S3-S10.
[31] Lita LV, Yu S, Niculescu S, et al. Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II; 2008. p. 877–82.
[32] Perotte A, Pivovarov R, Natarajan K, et al. Diagnosis code assignment: models and evaluation metrics. J Am Med Inf Assoc 2013;21(2):231–7. doi:10.1136/amiajnl-2013-002159.
[33] Koopman B, Zuccon G, Nguyen A, et al. Automatic icd-10 classification of cancers from free-text death certificates. Int J Med Inform 2015;84(11):956–65. doi:10.1016/j.ijmedinf.2015.08.004.
[34] Kang N, Singh B, Afzal Z, et al. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inf Assoc 2012;20(5):876–81. doi:10.1136/amiajnl-2012-001173.
[35] Elyne S, Boris C, Kim L, et al. Selecting relevant features from the electronic health record for clinical code prediction. J Biomed Inform 2017;74:92–103. doi:10.1016/j.jbi.2017.09.004.
[36] Suominen H, Ginter F, Pyysalo S, et al. Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description; 2007.
[37] Marafino BJ, Davies JM, Bardach NS, et al. N-Gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. J Am Med Inform Assoc 2014;21(5):871–5. doi:10.1136/amiajnl-2014-002694.
[38] Cao P, Yan C, Fu X, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online: Association for Computational Linguistics; 2020. p. 294–301.
[39] Ji S, Cambria E, Marttinen P. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. p. 73–8.
[40] Sadoughi N, Finley GP, Fone J, et al. Medical code prediction with multi-view convolution and description-regularized label-dependent attention. arxiv 2018.
[41] Falis M, Pajak M, Lisowska A, et al. Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Hong Kong: Association for Computational Linguistics; 2019. p. 168–77.
[42] Li F, Yu H. ICD coding from clinical text using multi-filter residual convolutional neural network. Proc AAAI Conf Artif Intell 2020;34(5):8180–7. doi:10.1609/aaai.v34i05.6331.
[43] Mayya V, S SK, Krishnan GS, et al. Multi-channel, convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries. Future Generat Comput Syst 2021;118:374–91. doi:10.1016/j.future.2021.01.013.
[44] He K, Zhang X, Ren S, et al. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–8.
[45] Shi H, Xie P, Hu Z, et al. Towards automated ICD coding using deep learning. arXiv preprint arXiv:171104075 2017.
[46] Baumel T, Nassour-Kassis J, Cohen R, et al. Proceedings of the Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence; 2018. p. 409–16.
[47] Guo D, Duan G, Yu Y, et al. A disease inference method based on symptom extraction and bidirectional long short term memory networks. Methods 2020;173:75–82. doi:10.1016/j.ymeth.2019.07.009.
[48] Rios A, Kavuluru R. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018; 2018. p. 3132–42.
[49] Du Y, Xu T, Ma J, et al. An automatic icd coding method for clinical records based on deep neural network. Big Data Res 2020;6(5):0. doi:10.11959/j.issn.2096-0271.2020040.
[50] Wang W, Xu H, Gan Z, et al. The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020; The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020; The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA. AAAI Press; 2020. p. 979–88.

[51] Gao Y, Fu X, Liu X, et al. Proceedings of 2021 IEEE International Conference on Bioinformatics and Biomedicine 2021.
[52] Teng F, Yang W, Chen L, et al. Explainable prediction of medical codes with knowledge graphs. Front Bioeng Biotechnol 2020;8:867. doi:10.3389/fbioe.2020.00867.
[53] Chelladurai J, Santhiappan S, Ravindran B. Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing; 2021. p. 225–37.
[54] Wang S, Ren P, Chen Z, et al. Few-shot electronic health record coding through graph contrastive learning. arXiv 2021.
[55] Devlin J, Chang MW, Lee K, et al. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–86. doi:10.18653/v1/N19-1423.
[56] Lee J, Yoon W, Kim S, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019;36(4):1234–40. doi:10.1093/bioinformatics/btz682.
[57] Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. 2020. arXiv:1904.05342.
[58] Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2022;3(1):1–23. doi:10.1145/3458754.
[59] Zhang Z, Liu J, Razavian N. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. p. 24–34.
[60] Pascual D, Luck S, Wattenhofer R. Proceedings of the 20th Workshop on Biomedical Language Processing. Online: Association for Computational Linguistics; 2021. p. 54–63.
[61] Luo J, Xiao C, Glass L, et al. Findings of the Association for Computational Linguistics: ACL-IJCNLP. Association for Computational Linguistics; 2021. p. 2096–101. doi:10.18653/v1/2021.findings-acl.184.
[62] Zhou T, Cao P, Chen Y, et al. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics; 2021. p. 5948–57. doi:10.18653/v1/2021.acl-long.463.
[63] Vu T, Nguyen DQ, Nguyen A. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence 2020. doi:10.24963/ijcai.2020/461.
[64] Xie X, Xiong Y, Yu PS, et al. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York, NY, USA: Association for Computing Machinery; 2019. p. 649–58.
[65] Wu Y, Zeng M, Fei Z, et al. Kaicd: a knowledge attention-based deep learning framework for automatic icd coding. Neurocomputing 2020. doi:10.1016/j.neucom.2020.05.115.
[66] Johnson A, Pollard T, Shen L, et al. MIMIC-III, A freely accessible critical care database. Sci Data 2016;3:16–35.
[67] Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV (version 1.0). 2021. doi:10.13026/s6n6-xd98.
[68] Pestian JP, Brew C, Matykiewicz P, et al. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. USA: Association for Computational Linguistics; 2007. p. 97–104.
[69] Goeuriot L, Kelly L, Suominen H, et al. In: Kelly L, Goeuriot L, Mandl T, editors Experimental IR Meets Multilinguality, Multimodality, and Interaction. Cham: Springer International Publishing; 2017. p. 291–303.
[70] Neves M, Butzke D, Dörendahl A, et al. Non-technical summaries of animal experiments indexed with icd-10 codes (version 1.0). 2019. Available from https://www.openagrar.de/receive/openagrar_mods_00046540.
[71] Goeuriot L, Suominen H, Kelly L, et al. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Cham: Springer International Publishing; 2020. p. 255–71.
[72] Subotin M, Davis AR. A method for modeling co-occurrence propensity of clinical codes with application to ICD-10-PCS auto-coding. J Am Med Inform Assoc 2016;23(5):866–71. doi:10.1093/jamia/ocv201.
[73] Pengtao X, Haoran S, Ming Z, et al. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018. p. 1066–76.
[74] Tsai SC, Huang CW, Chen YN. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2021. p. 4043–52. doi:10.18653/v1/2021.naacl-main.318.
[75] Rubin TN, Chambers A, Smyth P, et al. Statistical topic models for multi-label document classification. Mach Learn 2012;88:157–208.
[76] Bai T, Vucetic S. The World Wide Web Conference; 2019. p. 72–82.
[77] Tsai SC, Chang TY, Chen YN. Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). Hong Kong: Association for Computational Linguistics; 2019. p. 39–43. doi:10.18653/v1/D19-6206.
[78] Wang K, Chen X, Chen N, Chen T. Automatic emergency diagnosis with knowledge-based tree decoding. In: Bessiere C, editor. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. International Joint Conferences on Artificial Intelligence Organization; 2020. p. 3407–14.
[79] Zeng M, Li M, Fei Z, Yu Y, Pan Y, Wang J. Automatic ICD-9 coding via deep transfer learning. Neurocomputing. Deep Learning for Biological/Clinical Data 2019;324:43–50.
[80] Duque A, Fabregat H, Araujo L, et al. A keyphrase-based approach for interpretable icd-10 code classification of spanish medical reports. Artif Intell Med 2021;121:102177. doi:10.1016/j.artmed.2021.102177.