Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(7): 13379–13397.
DOI: 10.3934/mbe.2023596
Received: 31 March 2023
Revised: 12 May 2023
Accepted: 24 May 2023
Published: 09 June 2023

Research article

Construction of cardiovascular information extraction corpus based on electronic medical records

Hongyang Chang[1], Hongying Zan[1,2,*], Shuai Zhang[1], Bingfei Zhao[1] and Kunli Zhang[1,2]

[1] School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China
[2] Peng Cheng Laboratory, Shenzhen, China

* **Correspondence:** Email: iehyzan@zzu.edu.cn.

Abstract: Cardiovascular disease has a significant impact on both society and patients, making it necessary to conduct knowledge-based research such as research that utilizes knowledge graphs and automated question answering. However, the existing research on corpus construction for cardiovascular disease is relatively limited, which has hindered further knowledge-based research on this disease. Electronic medical records contain patient data that span the entire diagnosis and treatment process and include a large amount of reliable medical information. Therefore, we collected electronic medical record data related to cardiovascular disease, combined the data with relevant work experience and developed a standard for labeling cardiovascular electronic medical record entities and entity relations. By building a sentence-level labeling result dictionary through the use of a rule-based semi-automatic method, a cardiovascular electronic medical record entity and entity relationship labeling corpus (CVDEMRC) was constructed. The CVDEMRC contains 7691 entities and 11,185 entity relation triples, and the results of consistency examination were 93.51% and 84.02% for entities and entity-relationship annotations, respectively, demonstrating good consistency results. The CVDEMRC constructed in this study is expected to provide a database for information extraction research related to cardiovascular diseases.

Keywords: cardiovascular disease; corpus construction; electronic medical record

1. Introduction

According to the statistics released by National Center for Cardiovascular Diseases [1], cardiovascular disease (CVD) was the leading cause of death in China's urban and rural areas from the period of 2005 to 2019. With the change in modern people's living habits, the fatality rate of CVD is on the rise, with a trend of if occurring at younger ages. CVD has placed a huge burden on the social economy and citizens' health; it is therefore necessary to popularize the knowledge of CVD, such as its prevention, identification and treatment, and to reduce the difficulty of acquiring the relevant information, which

can reduce, to a certain extent, the harm and loss brought by CVD to the patients who suffer from it.

With the introduction of a series of specifications related to electronic medical records, such as the Specifications for Sharing Documents from Electronic Medical Record (EMR) *, the standardization and authenticity of EMRs have been recognized by the public and experts in the field of medical informatization. Therefore, EMRs have gradually become one of the main data sources for clinical medical research. EMRs record all of the information on patients from admission to discharge, as well as a lot of reliable clinical information, including examination indicators, treatment methods, changes in patients' vital signs, doctor's advice and so on. For example, "CT: post-coronary artery bypass grafting" reflects the diagnosis of the patient's post-coronary artery bypass disease through CT examination. The information contained in the EMRs gives a true picture of the patient's physical condition and the effectiveness of the treatment for the disease. The extraction of this information from the EMR text is in line with the requirements of informatization construction, with EMRs as the core, which can promote the informatization process of the medical industry.

The number of electronic medical records has increased by leaps and bounds since the implementation of the Health China Initiative and the construction of electronic medical record informatization. Manually extracting information from massive amounts of medical records is not only costly, but it also cannot keep up with the rate at which medical records are updated. As a result, current research is focused on how to effectively extract key information from massive amounts of data. Named entity recognition and relation extraction are two important branches of the information extraction task. Named entity recognition is the identification of entities with practical meaning from the given text, which, in medical texts, usually corresponds to disease, symptom, examination, body part, surgery, medication, etc.; relation extraction is the extraction of structured information from unstructured or semi-structured texts, and it represents the extracted data in the form of knowledge triples, e.g., in the above example "CT: post-coronary artery bypass grafting", the entities "CT", "post-coronary artery bypass grafting" and the relationship between the two entities "examination reveals disease" need to be extracted and expressed as <CT, examination reveals disease, post-coronary artery bypass grafting>. Due to the specialized nature of medical texts, there is currently a lack of a publicly available Chinese electronic medical record corpus for CVD, which, to a certain extent, impedes knowledge-based research on CVD. To this end, a Cardiovascular Disease Electronic Medical Record Entity and Entity Relationship Annotation Corpus (CVDEMRC) was constructed in this study in an attempt to provide a data basis for related research on information extraction, automated question and answer and teh intelligent diagnosis of CVD.

Focusing on the CVD electronic medical record text, this study mainly focuses on how to identify the entities present in the text and the relationship between the entities, suggests a standard system for their annotation and involved the construction of the CVDEMRC based on the system.

2. Related works

Many scholars in China and abroad are currently focusing on corpus construction, including the well-known medical reference terminology databases SNOMED RT [2], Clinical Terms Version 3 [3] and the modern Western medicine clinical standard terminology collection SNOMED CT [4], which combines the former two and further extends and updates them. In 2006, based on medical records

*http:// www.nhc.gov.cn/ mohwsbwstjxxzx/ s8553/201609/ d908a87908824fb8b4d42cba1b25dd3c.shtml

such as course records and discharge summaries, Meystre and Haug [5] constructed a named entity annotation corpus of 160 medical records, in which 80 common medical terms are involved. In 2008, Mayo Clinic scholars Savova et al. [6], for the first time, conducted a detailed classification of the modified information of entities and entity relationships and built a named entity corpus of disease entities based on 160 selected medical records. In 2009, in order to develop a system for the automatic extraction of important clinical information from medical records, Roberts et al. [7] worked on an annotated corpus of 20,000 cancer medical records and detailed it. The evaluation task of I2B2 in 2010 [8] suggested that teams participating in the evaluation could extract medical concepts, medical problems and the modification of problems from electronic medical records, and be able to identify the relationships that exist between medical problems, treatments and examinations. In addition to what has been mentioned above, some other researchers have done related work. In 2013, Morita et al. [9] used 50 fictitious electronic medical records in Japanese to construct an annotated corpus, which they utilized in the named entity recognition task of TCIR-10 MedNLP. Campillos et al. [10] constructed a corpus of named entities and entity relationships in the French language and did some other detailed work, including investigating the temporal information and relationships between medical events in electronic medical records.

In recent years, the work and research related to the construction of the Chinese medical corpus started relatively late, but it has made great progress. In 2013, Lei et al. [11] collected 800 electronic medical records from the Peking Union Medical College Hospital and constructed a named entity annotation corpus with annotation by two expert doctors. In 2014, Wang et al. [12] built a corpus of medical symptom names containing 11,613 chief complaints. In 2016, Yang et al. [13] constructed a Chinese electronic medical record named entity and entity relationship corpus based on 922 medical record texts. In 2019, Su et al. [14] constructed the first corpus of CVD risk factors in the field of Chinese health information processing. Zan et al. [15, 16] constructed an entity and relation annotation corpus for pediatric diseases, which contained 504 common diseases; in 2019, they built a knowledge base containing 8772 symptoms and 146,631 relations on the basis of the original named entity and relation annotation system. Guan et al. [17] constructed a Chinese medical information extraction dataset containing 11 types of entities and 44 types of entity relationships based on various data sources, such as textbooks and electronic medical records. In 2021, Ye et al. [18] collected the electronic medical records of diabetic patients and constructed a corpus of entities and relationships annotated for the EMRs of diabetes after several rounds of manual annotation.

The research on electronic medical records has also made significant progress, such as that by Wu et al. [19], who proposed an efficient and secure electronic medical record storage and query scheme; Beinecke et al. [20] used imaging data from medical records to predict the probability of prostate disease recurrence in patients through the use of machine learning strategies; Chang et al. [21] used a pre-trained model-based method to improve the results on both Chinese medical textbook data and medical record data; Hossain et al. [22] reviewed the mainstream methods used for studying electronic medical record data. At present, most of the information extraction research conducted for medical or electronic medical records is still based on supervised learning, and it is limited by the difficulty, cost and research direction of corpus construction, as well as the relevant available datasets being scarce; the CVDEMRC corpus built in this study can provide a data basis for many studies.

3. Establishment of entity and entity relationship annotation system for CVDEMRC

On the basis of the existing works, in terms of entity annotation, the study defines five types of entities, including disease, symptom, treatment (divided into three categories: surgical treatment, drug treatment and other treatment), examination and body. Attribute information about the entities of disease, symptom, treatment and examination are also involved, including modifier information, time information and numerical information (examination results). In terms of entity relationship annotation, the relationships are classified into six types based on entity annotation: disease and symptom, examination and disease, examination and symptom, treatment and disease, treatment and symptom, and body and symptom. In addition, there are six entity attribute relationships in attribute information: disease and modifier, symptom and modifier, treatment and modifier, disease and time, symptom and time, and examination and examination result.

3.1. Disease entity, symptom entity and their relationship

The disease is a diagnosis made by doctors according to patients' conditions. Specifically, it refers to the process of abnormal life activities caused by a disturbance in the autoregulation of the body after the onset of (symptoms of) the disease. The definition of a disease entity is mostly based on the International Classification of Diseases (ICD-10) and the section encoded as C (Diseases) in the Medical Subject Heading (MeSH). In addition to this, Baidu Encyclopedia and Medical Encyclopedia were used to help with identifying the concept of disease entities. Some examples of disease entities are as follows:

- *No history of diabetes, cerebrovascular disease (history of diabetes // cerebrovascular disease)*
- *No history of surgery, history of trauma, blood transfusion (history of trauma // blood transfusion)*
- *Cardiac ultrasound: postoperative aortic coarctation (post-operative aortic coarctation)*
- *Discharge diagnosis: 1) coronary heart disease; 2) erosive gastritis (coronary artery disease// erosive gastritis)*
- *Preoperative diagnosis: chest pain to be checked (chest pain to be checked)*

Symptoms are broadly defined as discomfort or abnormal reactions of the patient's body, and abnormal conditions are detected by examination. As there is no clear classification of symptoms in ICD-10, the Symptom Knowledge Base in Chinese [23] and diagnostics were referenced in the process of annotation for symptom entities. The symptom entities mainly include self-reported abnormalities by the patients themselves or by their families, as well as abnormalities detected by physicians through observation and medical imaging equipment. Here are some examples of typical symptom entity annotations:

- *Six months ago precordial discomfort without apparent cause, no chest tightness, panic, dizziness, headache, sweating (precordial discomfort// chest tightness// panic// dizziness// headache// sweating)*
- *Physical examination: no abnormal elevation in the precordial region (abnormal elevation)*
- *Mild mitral and tricuspid regurgitation (regurgitation)*

The relationship between the disease entity and the symptom entity is described as follows: disease causes symptoms (DCS). Examples are as follows:

- *Patient with recurrent chest pain with symptoms typical of acute myocardial ischemia <Acute myocardial ischemia, DCS, chest pain >*

- *Cough and sputum with chest tightness for more than 1 month after coronary artery bypass grafting for more than 2 months <Post-coronary artery bypass grafting, DCS, cough >, <Post-coronary artery bypass grafting, DCS, sputum >, <Post-coronary artery bypass grafting, DCS, chest tightness >*

3.2. The treatment entity and its relationships with the disease entity and the symptom entity

Treatment refers to the procedures, medications and interventions administered to patients to improve the autoregulation disorder, eliminate the cause of the disease or alleviate the symptoms. The treatment can be more finely divided according to the definition and means and methods of treatment; therefore, in this study, the entity 'treatment' is not annotated separately, but is divided into drug treatment, surgical treatment and other treatment.

Drugs can be broadly defined as a chemical substance capable of influencing the physiological functions or metabolic activities of the body. The drug treatment entities, in the process of annotation, are defined as those drugs that are coded D (Chemicals and Drugs) in the MeSH, and those drugs that are specified in the medical record as having been used by the patient or appearing in the medication instruction section. An example is as follows:

- *Medication guide rivaroxaban tablets (Bactrim 15 mg) 15 mg; (Rivaroxaban tablets // Bactrim)*

Surgical treatment refers to the process of cutting, incising or suturing a patient's body with a medical device such as a needle, knife or scissors for the purpose of maintaining the patient's health. In the process of annotation, surgical treatment entity is defined as the surgical concept coded as E4 (Surgical, procedures, Operative) in MeSH, and those surgeries to treat patients are conducted through some specific procedure indicated in the medical records. It is important to note that the "postoperative" category should be classified as disease. Some examples are as follows:

- *Had a cystectomy for a cyst 7 years ago and has recovered well since (cystectomy)*
- *Note sample: current diagnosis: postoperative adenocarcinoma of the right lung (postoperative adenocarcinoma of the right lung should be annotated as a disease)*

Other treatments mainly include radiation therapy, adjuvant therapy, chemotherapy and other types of treatment that are designed to fulfill certain therapeutic purposes but cannot be defined as drug treatment or surgical treatment, such as nutrition for the nervous system, free radical scavenging and circulation improvement. A typical example of other treatment entities is as follows:

- *Give symptomatic treatment such as, antiemetics, gastric protection and myocardial nutrition (antiemetics// gastric protection// myocardial nutrition)*

The relationships between treatment entity and disease/symptom entity are shown in Figure 1.

3.3. Examination entity and the relationships between examination entity and disease/symptom entity

Examination refers to the examination items, means and procedures carried out through the use of specific technology and medical equipment in order to confirm whether a patient is suffering from a certain disease or has certain symptoms, so as to provide a basis for doctors' clinical diagnosis and treatment. The definition of examination entity is referred to the medical imaging book [24]. In addition, minor examination items, physiological indicators, body fluid examination and observation by doctors

Entity 1	Entity Relation	Entity 2	Definition	Example
Treatment (include Drug treatment, Surgical treatment, Other treatment)	TID/TIS	Diseases/ Symptoms	Clearly indicate that the disease/ symptoms have improved after treatment	Positive for brucellosis, treated with rifamycin and doxycycline, symptoms improved significantly <Rifamycin, TID, brucellosis> <Doxycycline, TID, brucellosis>
	TWD/TWS		Treatment of worsening or failing to improve disease / Symptoms	Self-administered laxatives for constipation, later worsened walking difficulties <Laxatives, TWD, constipation>
	TLD/TLS		Diseases/symptoms resulting from the application of treatment	Chest discomfort after coronary artery bypass graft 1 year ago in our hospital <Coronary artery bypass graft, TLS, chest discomfort>
	TAD/TAS		Treatment given to disease/symptoms without mentioning effect	Initially diagnosed as post-operative adenocarcinoma of the left lung and given chemotherapy with "pemetrexed + carboplatin" <Pemetrexed, TAD, post-operative right lung adenocarcinoma> <Carboplatin, TAD, post-operative right lung adenocarcinoma>

TID/TIS: Treatment improved the disease/symptom.

TWD/TWDS: Treatment worsened the disease/symptom.

TLD/TLS: Treatment led to disease/symptom.

TAD/TAS: Treatment is applied to disease/symptom.

Figure 1. The relationships between treatment entity and disease/symptom entity and examples.

without the aid of instruments, such as palpation and auscultation, are also marked as examination entities. The following are some examples of examination entities:

- *Ambulatory ECG: Basal rhythm is sinus rhythm (ambulatory ECG)*
- *Body temperature 36.2 °C, pulse 71 beats/min (temperature // pulse)*

The relationship between the examination entity and the disease entity and examples are as follows:
1) Examination confirms disease (ECD)

- *Ultrasound suggests thyroid nodule; <Ultrasound, ECD, thyroid nodule >*

2) Examination is applied to confirm disease (ETCD)

- *No cancer seen in the off section. EGFR or NGS testing is recommended for No.2. wax block. <EGFR, ETCD, cancer >, <NGS, ETCD, cancer >*

The relationship between the examination entity and the symptom entity and examples are as follows:
3) Examination confirms symptom (ECS)

- *Chest DR shows: enlarged heart shadow <chest DR, ECS, enlarged heart shadow >*

4) Examination is applied for symptoms (EFS)

- *Atrial premature beats are seen during exercise. Coronary angiography is recommended. <Coronary angiography, EFS, premature atrial contractions >*
- *Bilateral mineral deposits in the substantia nigra; please consult with brain MRI and clinical history; <brain MRI, EFS, bilateral mineral deposits in the substantia nigra >*

3.4. Entity attribute

Attribute information in the annotation process is divided into modifier information for entities such as disease entity and symptom entity, time information for the disease entity and symptom entity, and numerical information on examination results.

3.4.1. Body information for entities

The body includes organs, body tissues, body systems, etc. But, in fact, there is no body entity in the Chinese EMR Named Entity and Entity Relationship Annotation Specification [25]. According to the analysis of cardiovascular electronic medical records, it is found that there is a large number of body parts and corresponding symptoms that are separated by modifying information in the medical record text, and one body part or modifier, etc., may correspond to multiple symptoms, e.g., in the record "symmetrical thorax, no localized augmentation, collapse, or pressure pain", where thorax is the subject of the symptoms, including symmetrical, localized augmentation, collapse and pressure pain. However, the word "no", which is used as a negative modifier, isolates the subsequent symptoms related to the thorax. If "symmetrical thorax" and the following symptoms were separately annotated as symptom entities, this would result in a lack of subject matter for the subsequent symptom entities, resulting in a large amount of missing information and affecting the authenticity of the EMR. To solve such problems, body entities have been added to the annotation process. The following are some typical examples:

- *Multiple abnormal signals in **bilateral thalamus** and **pontine brain**;*
- *No **thoracic** abnormalities, bilateral symmetry;*
- *No significant pathological murmur was heard in **any valve of the heart**;*

3.4.2. Modifier information of entities

Some qualitative or non-numerical descriptions of entities such as disease, symptom and other entities in EMRs reflect the relationship between entities and patients. This kind of information can be defined as the modifier information for entities, which can be, specifically, the subject, degree, frequency and state of the entity. This modifier information is important for doctors to grasp the patient's condition, and for the audience to accurately understand the contents of the medical record. Therefore, we have also annotated the modifier information in the annotation process, hoping to facilitate the progress of the research on EMR information extraction to a certain extent. The following is a list of the categories and examples of modifier information. The modifier information for disease entity and symptom entity is divided into eight categories: denial, non-patient, nature, severity, conditional, probable, unconfirmed and occasional.

1) Denial: Self-reported by the patient or relayed by families to confirm that it did not occur to the patient.

- ***No** history of blood transfusion*

- ***No heard*** *significant pathological murmurs in any of the heart valves*

2) Non-patient: It occurs not to the patient but to the patient's family members.

- ***Father*** *died of "stomach bleeding"*
- *2* ***sisters*** *with diabetes*

3) Nature: To determine the state of disease or symptom.

- *Bilateral facial numbness of* ***a transient nature***
- *The patient presented with* ***episodes of*** *syncope and dark confusion for more than 2 days*
- *Acute cerebral infarction,* ***small artery occlusion type***

4) Severity: Impact on the patient.

- ***Moderate-severe*** *tricuspid regurgitation*
- *Pulmonary hypertension (****moderate****)*
- *Hypertension grade 3* ***very high risk***

5) Conditional: It needs to happen under certain circumstances.

- *Painful cutting sensation on the foot* ***when touched***
- *High blood sugar* ***after meals***

6) Probable: No clear judgment can be made on the basis of current information.

- *The reappearance of nocturnal twitching and shaking of limbs, epilepsy* ***not excluded***
- *Disorders of consciousness to be investigated: epilepsy****?*** *TIA****?****(?//?)*

7) Unconfirmed: It is possible that this will happen in the future.

- *The* ***risk of*** *malignant arrhythmias*
- ***Prevention of*** *bilateral lower limb deep vein thrombosis*

8) Occasional: Infrequent occurrence, distinct from the persistent diseases or symptoms.

- ***Occasional*** *premature ventricular contractions*
- *Patient still has* ***occasional*** *dizziness*
- *Complaint:* ***intermittent*** *palpitations for 3 years*

3.4.3. Time information for entity

A patient's history of disease or symptoms is a significant reference for their current health state and therapeutic therapy; disorders like hypertension and coronary artery bypass grafting have a clear unfavorable impact on the cardiovascular system. Accordingly, it is necessary to clarify the time attribute information for disease and symptom entities in the annotation, and to divide them into two categories: past and persistent.

1) Disease or symptom that was present in the patient at some time in the past and is no longer ongoing.

- *Sudden intermittent coughing and choking on water* ***half a month ago***

2) Disease or symptom is occurring in the patient right now, including earlier happenings that are still occurring.

- *The feeling of tightness in the neck for* ***several minutes***
- *Found elevated blood pressure for* ***1 day***
- *Complaint: intermittent palpitations for* ***3 years***

3.4.4. Result information for examination entity

During the annotation process, it was discovered that many examination results in medical records are not represented in textual form, as they exist as numerical results of the examination to reflect whether the patient's indicators were within the healthy range. Although the deep learning algorithms currently in use are not numerically sensitive and do not reach the level of logical operations, it is believed that these numerical results directly reflect the patient's condition and may be useful in future studies, so we introduced the result information for examination entities. The following are some typical examples:

- *Heart rate* ***73 beats/min***
- *Occult blood 1+; erythrocytes* ***20.00/ μ L****; bacteria* ***33.00/ μ L***
- ***Weakly positive*** *for occult blood immunoassay*

4. The construction of the CVDEMRC

The most important task in building a corpus is to develop a reasonable annotation specification, and to annotate the corpus strictly according to it. The corpus construction process is therefore divided into the preparatory work and the formal annotation. Figure 2 depicts the overall annotation process.

4.1. Preparation

4.1.1. Data preparation and processing

The original corpus for this study was selected from the electronic medical records of patients in a tertiary hospital in Henan province, which contained documents such as admission records, medical course records, patient assessments, informed documents, surgical records and four other records. Furthermore, some of the surgical patient records include details of major surgery and unplanned reoperation reports.

It was found that the discharge summaries and discharge orders in the admission record, progress notes and other records in the medical records generally contained information regarding the patient's treatment during their stay in hospital. The admission record provides the basic information about the patient at the time of admission, which is the patient's initial status. The course record consists of the initial course record and the ward record, which document the examination and treatment received by the patient during the admission, as well as changes in the patient's condition, which is the patient's evolution information. The discharge summary documents the patient's fundamental information at the time of discharge, which is the information on the patient's treatment ending during their stay; the discharge medical advice provides the medication instruction information. Medication administration instructions are included in the discharge instructions. These four types of records in medical records serve as the initial corpus.

Statistical analysis of the raw data revealed that some of the patient records collected from hospitals were incomplete, with individual documents missing, so the records were screened and 200 complete

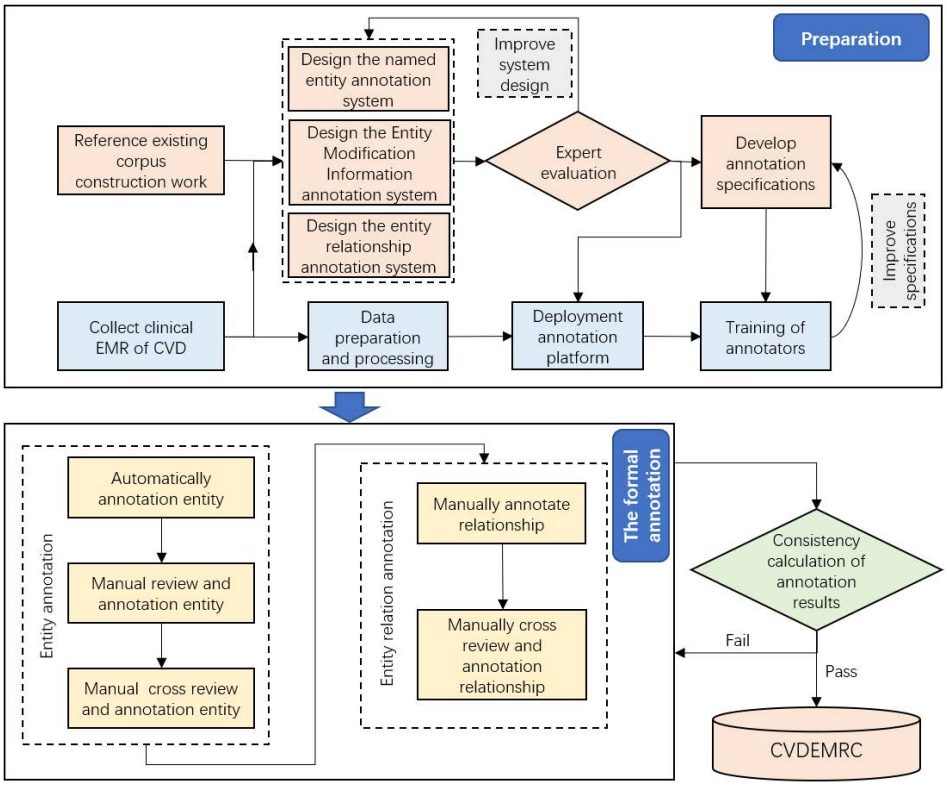


Figure 2. Flow chart of EMR annotation for CVD.

electronic medical records for CVD were selected finally. The screened data were desensitized prior to annotation, i.e., sensitive information was removed. Using the processing method used by Zhao et al. [26], sensitive information blocked in the I2B2-2006 [27] dataset was referenced, such as the patient's name, contact information, address and doctor's name. Here, we use a rule-based approach for processing.

4.1.2. The annotation platform development

The entity and entity relationship annotation platform developed by Zhang et al. [28] was changed for this research on the cardiovascular EMR annotation system and used for the annotation. By inputting the established annotation schema into the platform, the platform is able to automatically generate different colors based on category labels for the convenience of annotation personnel. For entity relation annotation, the platform utilizes a ray representation that connects the head and tail entities, and it annotates the relation label within the ray. In addition, the platform offers features such as word count statistics, annotation progress display and automatic annotation based on dictionary content and rules. These functions facilitate the control of annotation tasks and improve the efficiency of constructing corpora.

4.1.3. Training of annotators

Due to the high level of specialization in the medical field that is required for the annotation of electronic medical records, all annotation personnel were Master's students in computer science. In

order to ensure the correctness and consistency of annotations, as well as to familiarize the annotation personnel with the annotation platform, the system and the annotation guidelines, we allocated three electronic medical records for each annotator for trial annotation before officially annotating the corpus. In order to ensure the quality of annotation and avoid the impact of errors during the trial annotation process, we selected some data from the same source that were not included in the electronic medical records used to construct the corpus during the annotation phase, and we deployed it on the annotation platform. Additionally, during the trial annotation process, it is possible to conduct practical and in-depth observations on the collected medical record data, discover any omissions in the data during the formulation of the guidelines and supplement and improve the annotation guidelines based on these results.

4.2. The formal annotation

After data preparation, specification formulation, platform deployment and annotator training, a formal annotation of the corpus was carried out. Formal annotation is divided into three parts: semi-automatic annotation of entity, manual recheck of entity annotation and manual annotation of entity relationship.

Semi-automatic annotation of entity: It was discovered during pre-annotation that, while there were differences in the specific physical conditions of patients in the selected electronic medical records, the descriptions of the medical records mostly overlapped, particularly in the section on basic patient condition checking. To reduce labor costs and shorten the annotation time, medical records were divided according to specific punctuation marks. The specific operation is as follows: First, we segment the medical record document according to Chinese commas and periods to obtain a single isolated sentence; Then we summarize these sentences and keep only one duplicate sentence; Next, the annotator annotates these filtered sentences based on the sentence meaning in accordance with the specifications established above; Finally, the annotated results of the sentences will be used as templates, and rule-based methods will be used to automatically annotate all medical record texts.

Manual recheck of entity annotation: On the one hand, although the constructed sentence lexicon involves essentially all of the segmented sentence fragments covered in medical records, these sentence fragments could only express incomplete information and could not reflect the full context, which might lead to missing information and incorrect annotation results. On the other hand, the fact that medical records created by different doctors may contain multiple representations of the same condition also hampers the annotation process. As a result, after the automatic annotation of entities, the annotators need to manually check whether the automatic annotation is accurate, or whether any entities have been omitted in a sentence within the context. In the manual annotation stage, we used multiple rounds of the manual annotation strategy. First, annotator A checked the semi-automatic annotation results, filled in the missing entities and corrected the incorrect annotations and obtained the first round of annotation results. Then, annotator B rechecked and corrected the results based on the first round and recorded the inconsistent parts to obtain the second round of annotation results. Finally, they discussed and solved the doubts and issues, and A made modifications to obtain the final third round of annotation results.

Manual annotation of entity relationship: Semi-automatic annotation of entity improves efficiency, but the semi-automatic annotation method for electronic medical record text cannot support entity relationship annotation because the entity relationship annotation requires comprehensive consideration of the whole sentence, or even cross-sentence information. The procedure of entity relationship

annotation is roughly similar to that of manual rechecks described above. First, the annotator determines the relationship between these entities based on the entity annotation results, forming the first round of annotation results. Then, another annotator verified and improved the results of the second round of annotation and recorded any issues with conflicting opinions. After the final discussion and resolution, the third round of annotation results were obtained.

5. Statistical analysis

5.1. Calculate consistency

To evaluate the reliability of this annotation, we selected the first and third rounds of annotation results from the manual annotation stage and calculated the consistency between the two results. The following equations can be used to calculate consistency:

$$P = \frac{A_1 \cap A_2}{A_1} \tag{5.1}$$

$$R = \frac{A_1 \cap A_2}{A_2} \tag{5.2}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{5.3}$$

where A_1 and A_2 respectively represent the different annotation results of the two annotators, while $A_1 \cap A_2$ represents the consistent part of the annotation results between the two annotators; Equations (5.1) and (5.2) represent the proportion of consistently labeled results in two different results, respectively; Equation (5.3) is a classic formula for calculating F1 values, which can effectively balance the influence of two results.

Table 1. The consistency results for cardiovascular corpus annotation and the number of entities and entity relationships. N* indicates the number after removing duplicates.

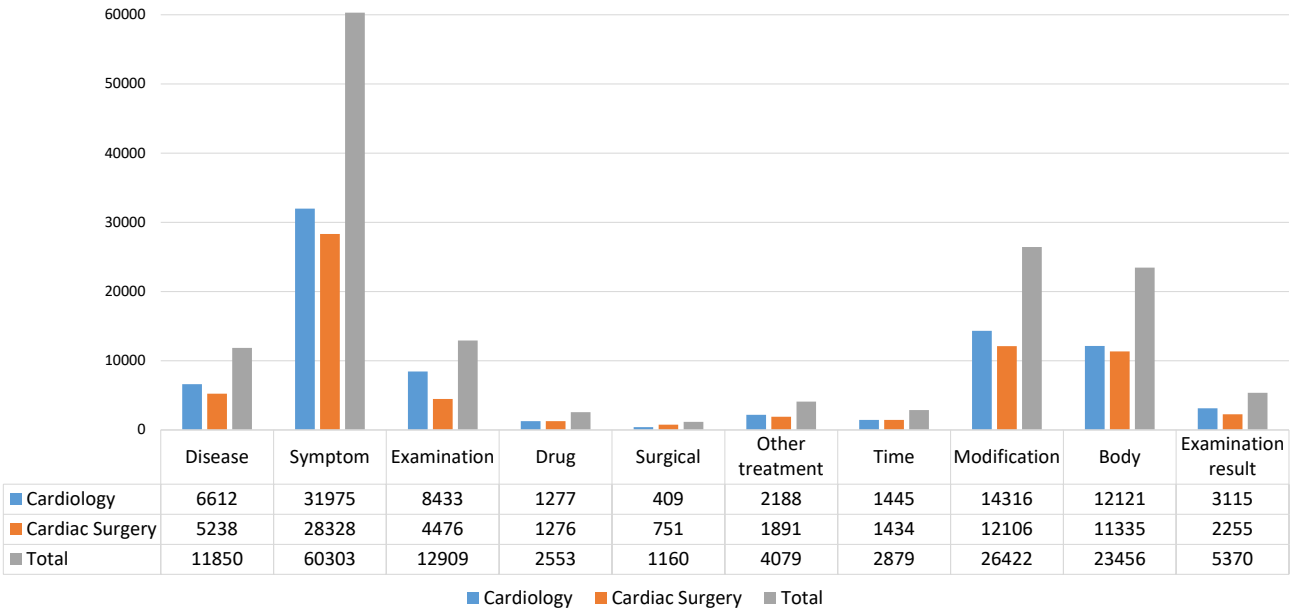
	Cardiology		Cardiac Surgery		CVD	
	entity	relationship	entity	relationship	entity	relationship
P (%)	89.01	80.27	96.48	85.88	93.20	83.15
R (%)	91.67	82.14	95.06	87.93	93.97	85.66
F (%)	90.32	81.19	95.76	86.89	93.58	84.39
Num*	5087	6927	4066	5144	7691	11,185
Num	81,891	39,447	69,090	37,367	150,981	76,814

The results of annotation consistency are shown in the upper part of Table 1. According to Artstein and Poesio [29], the calculated result of annotation consistency achieved 80%, indicating that the consistency of the annotation corpus is acceptable. The consistency rate of named entity annotation reached 93.51%, and that of entity relationship annotation reached 84.02. Additionally, Table 2 shows the number of entities and entity relationships included in the corpus in the lower half. The consistency demonstrates that the CVD medical record annotation corpus is acceptable. The high consistency of the

entities shows the effectiveness of employing sentence lexicon annotation and semi-automatic entity annotation, as proposed in this paper, for medical records data.

5.2. *Corpus statistics*

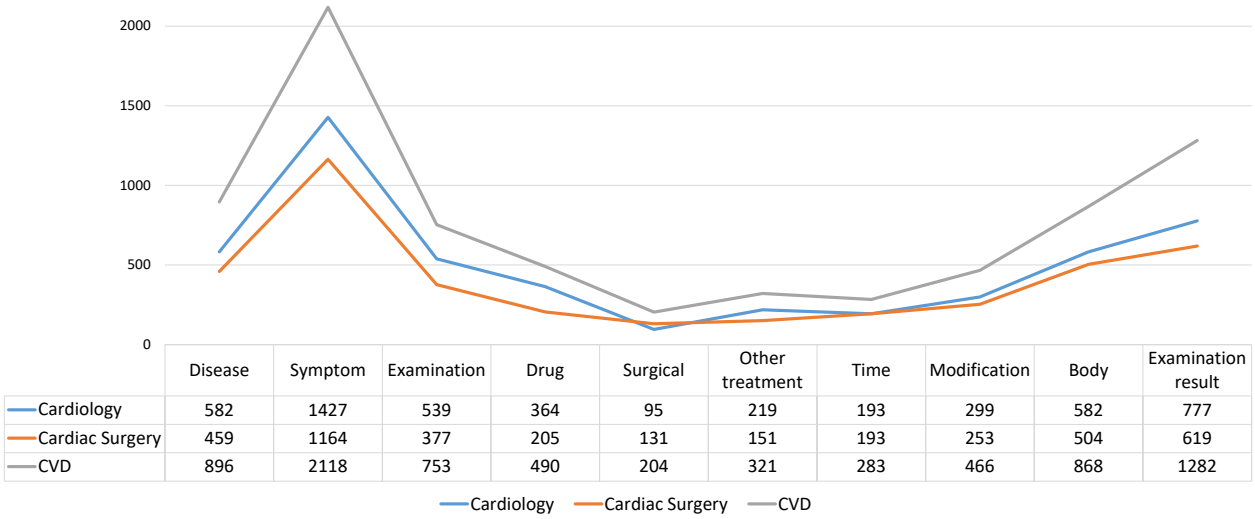
The annotation consistency of medical records from the cardiology department is significantly lower than that of the cardiac surgery department. As can be ascertained from the statistics of the number of different entities in Figure 3, most of the entities in the medical records of the cardiology department, such as the symptom entity, examination entity and disease entity, are more numerous than those in the cardiac surgery department, with the exception of the surgical treatment entity. This indicates that the information density of the cardiology department's EMRs is relatively high.



	Disease	Symptom	Examination	Drug	Surgical	Other treatment	Time	Modification	Body	Examination result
Cardiology	6612	31975	8433	1277	409	2188	1445	14316	12121	3115
Cardiac Surgery	5238	28328	4476	1276	751	1891	1434	12106	11335	2255
Total	11850	60303	12909	2553	1160	4079	2879	26422	23456	5370

Figure 3. Statistics on the number of different types of entities.

Furthermore, the statistical results of the numbers of entity categories are shown in Figure 4. It can be seen that the number of entities after deduplication is basically similar to that of non-deduplicated entities. In the medical records of the cardiology department, the number of deduplicated entities of most entity categories was higher than that of the cardiac surgery department, except for the surgical treatment entity, indicating that the diseases in the cardiology department were more diverse. The high information density, diversified symptoms and different expressions pose a significant challenge to the annotation process, resulting in low consistency in the annotation for the cardiology department EMRs. As entity relationship annotation is based on the results of entity annotation, the situation of relationship annotation is broadly consistent with that of entity annotation.



	Disease	Symptom	Examination	Drug	Surgical	Other treatment	Time	Modification	Body	Examination result
Cardiology	582	1427	539	364	95	219	193	299	582	777
Cardiac Surgery	459	1164	377	205	131	151	193	253	504	619
CVD	896	2118	753	490	204	321	283	466	868	1282

Figure 4. The statistical results regarding the number of different types of entities after removing duplicates.

5.3. Dataset statistical analysis

We segmented the annotation results of electronic medical records at the document level according to the needs of the entity relation extraction task, forming a sentence-level CVD entity relationship dataset, then filtered out the relationship categories with a quantity of less than 10 and conducted statistics and analysis on it. The processed dataset contained a total of 6077 corpora, with an average length of 72.23 characters and a total of 25,435 triples. The number of entities and attributes in the dataset after deduplication was 4932, forming 16 categories, 40 subcategories and 9293 triples.

Table 2. The statistical information on entities and entity attributes.

Entity	Num		Entity	Num	
	Deduplicate	All	attribute	Deduplicate	All
Disease	473	7230	Time	202	1748
Symptom	1349	16,225	Body	670	5813
Exam	475	5510	Modifier	293	8494
Surgical treat	71	204	Result	1001	2524
Drug treat	279	1475			
Other treat	119	1647			
Total	2766	32,291	Total	2166	18,579

The statistical information on entities and entity attributes is shown in Table 2. The statistical information on triples is shown in Table 3. The two types of entity relationships, "Symptom-Exam" and "Disease-Drug", had the highest numbers due to the presence of a large number of examination items in the medical records, including general examinations and examinations specific to diseases and symptoms. Most patients had other underlying diseases that require long-term medication. In the medication guidance section, doctors provide medication usage guidance based on the patient's

condition, so there is a significant relationship between medication treatment and the disease. In the entity attribute triplet, the location information and modifier information for symptoms had the highest number of attributes, and there was a large number of cases in medical records where there were multiple lesions or abnormalities in a certain location. The reason why there are many modifying attributes of symptom entities is that, in addition to modifying existing symptoms in medical records, there is a large amount of modifying information that negates symptoms in general examinations.

Table 3. The statistical information on triples.

Entity relation	Num		Entitie attribute	Num	
	Deduplicate	All	relation	Deduplicate	All
Disease-Symptom	368	544	Disease-Time	246	378
Disease-Exam	314	450	Disease-Modifier	327	2774
Disease-Surgical	128	171	Disease-Body	65	150
Disease-Drug	1035	1237	Symptom-Time	718	1370
Disease-Other	1004	1526	Symptom-Modifier	871	5720
Symptom-Exam	1082	2536	Symptom-Body	1592	5663
Symptom-Surgical	26	33	Exam-Result	1213	2524
Symptom-Drug	195	238			
Symptom-Other	109	121			
Total	4261	6856	Total	5032	18,579

In addition, we randomly divided the training set, validation set and test set in a ratio close to 8:1:1, each containing 4856, 614 and 607 corpora. The phenomenon of triple entity overlap and the inclusion of multiple triples in a single corpus are important issues that need to be addressed in entity relationship extraction tasks. These two issues are particularly prominent in Chinese medical data. Therefore, we conducted statistical analysis on the dataset for these two situations, and the statistical information is shown in Table 4. Triple entity overlap refers to the phenomenon of entity sharing between triples, which can be divided into three types: non-overlapping type (Normal), single-entity overlap type (SEO) and entity-pair overlap type (EPO). Normal type refers to entities in all triples included in the corpus that only participate in entity-pair matching once in this corpus, and entity pairs in triples have only one relationship in the entire corpus. SEO type refers to at least one entity participating in two or more entity-pair matches in a corpus. The type of EPO refers to the situation where there are multiple semantic relationships between an entity pair in a single corpus or multiple corpora.

Statistics on the number of triples in the corpus show that over 60% of the three subsets contained two or more triples, and over 25% contained five or more triples. The entity-overlap statistics of the corpus show that more than half of the corpus has entity overlap, and that one of the corpora may belong to both the SEO and EPO types. Due to the fact that EPO statistics include both single-sentence and cross-sentence corpora, the more corpora in the dataset, the higher the proportion of entity-pair overlap. Therefore, the proportion of EPO corpora in the training set was higher than that in the test and validation sets. Due to the frequent occurrence of multiple symptoms corresponding to a single examination or location in the medical record, the proportion of SEO corpora was correspondingly higher.

Table 4. The number of corpora containing different numbers of triples and different types of overlapping triples.

Type	Train		Validate		Test	
	Num	Pre (%)	Num	Pre (%)	Num	Pre (%)
# Number of corpora containing different numbers of triples #						
1	1687	34.74	221	35.99	213	35.09
2	862	17.75	110	17.92	106	17.46
3	487	10.03	60	9.77	69	11.37
4	506	10.42	61	9.93	65	10.71
>=5	1314	27.06	162	26.38	154	25.37
# Number of corpora with overlap triple of different types #						
Normal	1754	36.12	279	45.44	272	44.81
SEO	2665	54.88	326	53.09	327	53.87
EPO	1747	35.98	58	9.45	56	9.23
# Number of corpora	4856	79.91	614	10.1	607	9.99
# Number of triples	20,311	79.85	2659	10.45	2465	9.69

6. Conclusions

Based on the analysis of the electronic medical records of CVD, we developed the annotation system and specification in combination with numerous relevant studies, and we constructed the CVDEMRC using the semi-automatic annotation method with multiple rounds of recheck. The CVDEMRC contained a total of 7691 named entities (i.e., entities after deduplication) and 11,185 entity relationships. Both entity and entity relationship annotations have passed the consistency test, which can provide a data basis for further research on CVD.

Experience and summary The semi-automatic annotation method for the annotation of named entities in electronic medical records was proposed after the preliminary analysis. This method has improved the annotation efficiency and quality to some extent; however, it cannot be applied to entity relationship annotation for the time being. In future work, based on the existing medical record corpus, we will strive to perform the first round of pre-annotation by using deep learning algorithms to further reduce human labor costs and improve annotation efficiency. In addition, although Yang et al. [25] used an annotation group of six annotators, our annotation team comprised 20 computer science Master's students majoring in natural language processing. The increase in the number of annotators may have made it difficult to control the quality of annotation, so we chose annotators with experience in annotating textbooks on CVD to solve such a problem, along with performing adaptive pre-annotation, constructing sentence lexicon, cross-rechecking in multiple rounds, etc. As an attempt to use a larger group of annotators with low time cost to efficiently complete a more specialized annotation task in the field, it is hoped that this study will make some contributions to supplementing the gaps in the research on the Chinese medical information extraction corpus.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and gratefully acknowledge the support of the Major Projects of China National Social Science Fund (21&ZD338).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. N. Health, F. P. C. of the People's Republic of China, Electronic medical records application management standards (trial), *Chin. Pract. J. Rural Doctor*, **24** (2017), 3.
2. K. A. Spackman, K. E. Campbell, R. A. Côté, Snomed rt: a reference terminology for health care, in *Proceedings of the AMIA Annual Fall Symposium*, American Medical Informatics Association, (1997), 640.
3. M. O'neil, C. Payne, J. Read, Read codes version 3: a user led terminology, *Methods Inf. Med.*, **34** (1995), 187–192. https://doi.org/10.1055/s-0038-1634585
4. M. Q. Stearns, C. Price, K. A. Spackman, A. Y. Wang, Snomed clinical terms: overview of the development process and project status, in *Proceedings of the AMIA Symposium*, American Medical Informatics Association, (2001), 662.
5. S. Meystre, P. J. Haug, Natural language processing to extract medical problems from electronic clinical documents: performance evaluation, *J. Biomed. Inf.*, **39** (2006), 589–599. https://doi.org/10.1016/j.jbi.2005.11.004
6. G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, et al., Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, *J. Am. Med. Inf. Assoc.*, **17** (2010), 507–513. https://doi.org/10.1136/jamia.2009.001560
7. A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, et al., Building a semantically annotated corpus of clinical texts, *J. Biomed. Inf.*, **42** (2009), 950–966. https://doi.org/10.1016/j.jbi.2008.12.013
8. Ö. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text, *J. Am. Med. Inf. Assoc.*, **18** (2011), 552–556. https://doi.org/10.1136/amiajnl-2011-000203
9. M. Morita, Y. Kano, T. Ohkuma, M. Miyabe, E. Aramaki, Overview of the ntcir-10 mednlp task., in *NTCIR*, (2013), 1.
10. L. Campillos, L. Deléger, C. Grouin, T. Hamon, A. L. Ligozat, A. Névéol, A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot), *Lang. Resour. Eval.*, **52** (2018), 571–601. https://doi.org/10.1007/s10579-017-9382-y
11. J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu, A comprehensive study of named entity recognition in chinese clinical text, *J. Am. Med. Inf. Assoc.*, **21** (2014), 808–814. https://doi.org/10.1136/amiajnl-2013-002381

12. Y. Wang, Z. Yu, L. Chen, Y. Chen, Y. Liu, X. Hu, et al., Supervised methods for symptom name recognition in free-text clinical records of traditional chinese medicine: an empirical study, *J. Biomed. Inf.*, **47** (2014), 91–104. https://doi.org/10.1016/j.jbi.2013.09.008

13. J. Yang, Q. Yu, Y. Guan, Z. Jiang, An overview of research on electronic medical record oriented named entity recognition and entity relation extraction, *Acta Autom. Sin.*, **40** (2014), 1537–1562.

14. J. Su, B. He, H. Wu, J. Yang, Y. Guan, J. Jiang, et al., Cardiovascular disease risk factor labeling system and corpus construction based on Chinese electronic medical records, *Acta Autom. Sin.*, **45** (2019), 420. https://doi.org/10.16383/j.aas.2018.c170206.

15. H. Y. Zan, T. Liu, C. Y. Niu, Y. Zhao, Y. Zhang, Z. Sui, Construction and application of named entity and entity relations corpus for pediatric diseases, *J. Chin. Inf. Process.*, **34** (2020), 19–26.

16. H. Zan, Y. Han, Y. Fan, C. Niu, K. Zhang, Z. Sui, Construction and analysis of symptom knowledge base in chinese, *J. Chin. Inf. Process.*, **34** (2020), 33–40.

17. T. Guan, H. Zan, X. Zhou, H. Xu, K. Zhang, Cmeie: Construction and evaluation of Chinese medical information extraction dataset, in *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*, Springer, (2020), 270–282.

18. Y. Ye, B. Hu, K. Zhang, H. Zan, Construction of corpus for entity and relation annotation of diabetes electronic medical records, in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, (2021), 622–632.

19. Z. Wu, S. Xuan, J. Xie, C. Lin, C. Lu, How to ensure the confidentiality of electronic medical records on the cloud: A technical perspective, *Comput. Biol. Med.*, **147** (2022), 105726. https://doi.org/10.1016/j.compbiomed.2022.105726

20. J. M. Beinecke, P. Anders, T. Schurrat, D. Heider, M. Luster, D. Librizzi, et al., Evaluation of machine learning strategies for imaging confirmed prostate cancer recurrence prediction on electronic health records, *Comput. Biol. Med.*, **143** (2022), 105263. https://doi.org/10.1016/j.compbiomed.2022.105263

21. H. Chang, H. Zan, T. Guan, K. Zhang, Z. Sui, Application of cascade binary pointer tagging in joint entity and relation extraction of chinese medical text, *Math. Biosci. Eng.*, **19** (2022), 10656–10672. https://doi.org/10.3934/mbe.2022498

22. E. Hossain, R. Rana, N. Higgins, J. Soar, P. D. Barua, A. R. Pisani, et al., Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review, *Comput. Biol. Med.*, **155** (2023), 106649. https://doi.org/10.1016/j.compbiomed.2023.106649

23. H. Zan, Y. Han, Y. Fan, C. Niu, K. Zhang, Z. Sui, Establishment and analysis of chinese symptom knowledge base, *J. Chin. Inf. Process.*, **34** (2020), 30–37.

24. E. Wu, *Medical Imaging*, 5th edition, 2003.

25. J. Yang, Y. Guan, B. He, C. Qu, Q. Yu, Y. Liu, et al., Corpus construction for named entities and entity relations on Chinese electronic medical records, *J. Software*, **27** (2016), 2725–2746.

26. Y. S. Zhao, K. L. Zhang, H. C. Ma, K. Li, Leveraging text skeleton for de-identification of electronic medical records, *BMC Med. Inf. Decis. Making*, **18** (2018), 65–72. https://doi.org/10.1186/s12911-018-0598-6

27. O. Uzuner, P. Szolovits, I. Kohane, i2b2 workshop on natural language processing challenges for clinical records, in *Proceedings of the Fall Symposium of the American Medical Informatics Association*, Citeseer, 2006.

28. K. Zhang, X. Zhao, T. Guan, B. Shang, Y. Li, H. Zan, Construction and application of medical text oriented entity and relationship annotation platform, *J. Chin. Inf. Process.*, **34** (2020), 117–125.

29. R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, *Comput. Ling.*, **34** (2008), 555–596. https://doi.org/10.1162/coli.07-034-R2

AIMS Press