# Constructing a fine-grained entity recognition corpus based on clinical records of Traditional Chinese Medicine

**tingting zhang**
Chengdu University of Traditional Chinese Medicine https://orcid.org/0000-0002-1304-1149
**Yaqiang Wang**
Chengdu University of Information Technology
**Xiaofeng Wang**
Chengdu University of Information Technology
**Yafei Yang**
Chengdu University of Traditional Chinese Medicine
**Ying Ye** ( yeyingtcm@163.com )
https://orcid.org/0000-0001-9042-8008

**Research article**

**Keywords:** TCM clinical records, Fine-grained annotation, Corpus construction, Guideline development, Named entity recognition

**Posted Date:** October 24th, 2019

**DOI:** https://doi.org/10.21203/rs.2.16418/v1

# Abstract

Objective  In this paper, we focused on building a fine-grained entity annotation corpus with corresponding annotation guideline of Traditional Chinese Medicine ( TCM ) clinical records, and to provide an effective way to build more corpora of TCM clinical records in the future.

Methods: Instead of previous research methods, we proposed four steps approach which is suitable for TCM medical records in our corpus construction work. Firstly, determine the entity types included in this study through sample annotation method; secondly, draft our fine-grained annotation guideline by summarizing the characteristics of the dataset and referring to some existing guidelines; thirdly, update the guideline through iterative annotations way until the inter-annotator agreement (IAA) value exceed 0.9, kappa value was used to measure the IAA; fourthly, comprehensive annotations were performed, if IAA value exceeds 0.9 stably . After above four method steps, we succeeded to construct the fine-grained entity recognition corpus of TCM clinical records.

Results: There are 4 entity categories involving 13 entity types being determined finally. The fine-grained annotated entity corpus consists of 1104 entities and 67799 tokens totally. The final IAAs are 0.93, 0.94, 0.94 respectively (between two of the three annotators), the IAA value show the fine-grained entity recognition corpus are of high quality. We constructed a fine-grained annotated guideline and entity recognition corpus of TCM clinical records.

Conclusions: The four-step method was of high quality, the corpus constructed in this study was an encouraging example . Based on this approach, more comprehensive corpus about TCM clinical records will be built to support the TCM named entity recognition tasks in future research.


# 1. Introduction

Chinese Electronic Medical Record (EMR) contains much information about clinical diagnosis and treatment events. Since the implement of the “Basic Norms of Electronic Medical Records” in China and developing information technology, solid data foundation is generated as a result of unprecedented expansion of the EMRs. Electronically stored clinical documents may contain both structured data and unstructured data. Although the Chinese Ministry of Health (MOH) has issued a series of relevant regulations [1], most EMRs have much more clinical information stored as unstructured data in clinical narrative, such as the chief complaint, clinical decision making etc. Traditional Chinese medicine (TCM) is a unique and complicated medical system which has developed over thousands of years [2]. It is a complementary and alternative medical system in Western countries. With the continuous development of information technology, much of the knowledge in large-scale TCM clinical records has gradually been dug up in data-driven medical studies, clinical decision making, and health management.

Natural language processing (NLP) for TCM clinical medical records has become a hot topic in recent years, it promotes standardization which is necessary in automatically processing and analyzing EMRs, it is also widely believed to be the most useful method to improve the efficiency of biomedical text mining.

Named Entity Recognition (NER) [3,4] is a one of the high-level task of NLP.A human annotated entity corpus is indispensable resource for training and testing the performance of these automated systems of NER tasks. In English, some medical knowledge bases contribute to NER of clinical records, such as terminology systems like UMLS [5], clinical Ontology systems like SNOMED CT [6], and medical databases like DrugBank [7]. In China, some resources have been established for the development of NER tasks in TCM, for example, Traditional Chinese Medicine Language System (TCMLS) standardized the terminology definition of TCM. A certain number of entity types of Chinese clinical records has already been annotated, such as medication, anatomy, treatment, test, symptom, body part, temporal word, drugs, operation, etc. [4,8–12]. However, to the best of our knowledge, the open Chinese annotated corpora are rarely about TCM clinical records, and the methods of corpus construction are not standard and effective, for one reason, lack of the TCM clinical dataset. Due to concerns regarding patients' privacy as well as concerns about revealing unfavourable institutional practices [13], the records are very private and scarce. For another reason, high degree of difficulty. The Chinese clinical text has sublanguage features [14], the characteristics of raw TCM free-text clinical records are different a lot from the common texts which are noted by normal Chinese language, the characteristics are narrative form, concise and classical-Chinese-like style, nonstandard description [15]. Consequently, constructing a corpus of TCM clinical records is facing great difficulty and electronic capture or retrieval of TCM clinical texts data has been challenging, thus the NLP tasks about TCM clinical free text is still at a preliminary stage.

In general, it takes 4 steps to build a corpus in biological information domain, they are (1) data selection, (2) guideline drafting, (3) annotation, (4) consistency assessment, updating the guideline and reannotation. However, most of the relevant researches are concise and lack of detailed descriptions currently, especially in step (1), (2), and (3). There is no existing fine-grained annotation schema applicable in TCM clinical domain. In our study, we followed this standard approach and proposed an improved four-step method to make the process clear and replicable. This paper has six parts, of which the main contents are as the following parts: the related work are summarized in section 2; then we describe why and how to choose the dataset and entity types in our study in section 3; next, we narrate the development of annotation guideline, annotation method, and consistence assessment in section 4; in section 5, we presents the inter-annotator agreement (IAA) values, result analysis of the annotations; finally in section 6, we summarize the contribution of this study and look forward to the future work.

Table 1 Related researches about corpus construction of Chinese clinical text in recent 5 years.

| year | author | scale and object | entities | fine-grained | TCM clinical texts |
|---|---|---|---|---|---|
| 2014 | Xu et al.[8] | 336 Chinese discharge summaries of 71 355 words | Medication, anatomy, medical problem, treatment, test | N | N |
| 2014 | Lei et al.[4] | 400 admission notes and 400 discharge summaries | Clinical problems, procedures, laboratory test, medications | N | N |
| 2014 | Wang et al. [16] | 11613 clinical records | symptoms | N | Y |
| 2014 | Wang et al. [17] | 115 electronic medical records | 115 documents tumor-related information from operation notes of hepatic carcinomas | N | N |
| 2014 | Gao et al.[18] | 42 health records of stroke | body structures and clinical description | N | Y |
| 2015 | Li et al.[19] | 700 initial diagnosis records, CHF data of 253 cases. | TCM herb, Symptom | N | Y |
| 2015 | Xu et al.[20] | 24817 de-identified Chinese EMRs | Symptom, clinical test, disease, drug, body part, and procedure categories | N | Y |
| 2016 | Zhang et al. [21] | 2000 notes (1,000 admission notes and 1,000 discharge summaries) | Disease and syndrome, symptom and sign, treatment and drug, and laboratory test | N | N |
| 2016 | Wan et al. [22] | more than 100 000 TCM article abstracts | Herb, syndrome, disease, formula | N | Y |
| 2016 | Liu et al.[12] | 1778 clinical notes of 281 hospitalized patients | Temporal expression (TE) and normalization in Chinese clinical notes (type, value and modifier) | N | N |
| 2017 | Ruan et al. [23] | 1000 EMRs | Symptoms, departments, diseases, medicines, and examinations | N | Y |

| | | | | | |
|---|---|---|---|---|---|
| 2017 | He et al.[9] | 500 discharge summaries and 492 progress notes | Diseases, symptoms, treatments | N | N |
| 2018 | Zhang et al.[24] | 400 documents | Symptom, test, diagnosis, treatment, and body part | N | N |
| 2018 | Miao et al.[25] | 540 reports | Breast Imaging Reporting and Data System (BI-RADS) | N | N |
| 2018 | Bao et al.[26] | 600 documents | The history of present illness, past history, personal history, family history | N | N |
| 2019 | Wang et al.[27] | 1596 annotated instances (10,024 sentences) | Diseases, symptoms, exams, treatments, and body parts | N | N |
| 2019 | Gao et al.[10] | 255 authentic admission records | Medical discovery, Body part, Temporal word, Disease, Medication, Treatment, Inspection, Laboratory test and Measurement. | N | N |
| 2019 | Cai et al.[11] | 1000 admission records | Anatomical Part, Symptom Description, Independent Symptoms, Drug, Operation | N | N |
| 2019 | Xiong et al.[28] | 1000 admission notes and 800 discharge summaries | body, disease, symptom, test and treatment | Y | N |

## 2. Related Work

Recently, the clinical EMRs research has become a hot topic[29]. The researches of English EMR entity corpus starts early, and text mining and NLP applications, algorithms, and corpus based on English language has been relatively mature. There are some well-known publicly available annotated corpora, such as the GENIA[30] for data mining and information extraction in molecular biology domain, CADEC[31] for adverse drug events, NCBI Disease[32] for disease names and adverse effect, DDI[33] for pharmacological substances and drug–drug interactions. Moreover, the Integrating Biology and the Bedside(i2b2) challenges contributed to the clinical NLP researches. I2b2 has organized a medical information extraction challenge from the English discharge summaries in 2009 and 2010, the concepts of extraction involve drugs, doses, duration, medical problems, treatment, testing, etc. From 2006 to now, i2b2 has released 9 corpora on evaluating EMRs information extraction. Based on these corpora, a lot of progress on NER researches on English discharged summaries come out. Currently, the availability of a large corpus in Chinese is limited, and the development of corpus construction in Chinese medicine has

fallen behind the progress of English of Western medicine. The annotation scheme and evaluation method of corpus construction in English are of great high reference value for Chinese clinical notes. Influenced by the rapidly developing English medical corpus, the Chinese medical corpus has gradually begun to develop. The major annotated corporaof Chinese medical notes are summarized in *Table 1* and described detailed in the following part.

### 2.1 Clinical entity recognition corpus construction

Based on the concept annotation guideline in 2010 I2B2 challenge, Xu et al.[8] has labeled a set of standard corpus in 336 Chinese discharge summaries (Medication, anatomy, medical problem, treatment, test) in 2014. The annotation work contains two rounds, the first round was made by three annotators with relevant domain background, the second round was conducted by three annotators with backgrounds in computer linguistics. Refined results and a final gold standard were obtained by combining the results from the above two rounds of annotation. Lei et al[4] constructed an annotated entity corpus of 400 discharge summaries and 400 admission notes. The guideline was similar to those used in the 2010 i2b2 NLP challenge, but differently, the “treatments” was divided into “procedures” and “medications”. Moreover, Wang et al[17] annotated 12 elements doctors wanted to get from a free-text operation note, in this study, the guideline was not mentioned and the annotation process is briefly described. Miao et al[25] annotated BI-RADS categories manually. It is really a first study on information extraction from Chinese breast ultrasound reports. These two studies of Wang et al and Miao et al are good examples of information extraction for particular information. Liu et al[12] annotated TEs in the clinical notes by the annotation guidelines which refereed to the TE annotation guidelines of TimeML for English newswire text and the 2012 i2b2 NLP challenge for English clinical text. Furthermore, in 2019, based on Resident Admit Notes (RANs), Gao et al.[10] described a more detailed method of constructing a corpus of nine entity types. The guideline was also developed and refined on the bases of i2b2 annotation guidelines, differently, they added “body part” entity and “temporal word” entity in their annotation work, and “inspection” entity and “laboratory test” entity is distinguished. An iterative annotation method was put into practice to form the manual annotation scheme. Furthermore, He et al[9] referred the annotation method in English clinical text, and built a syntactic corpus about entity diseases, symptoms, treatments. They built the draft guideline first, then trained the annotators and updating the guideline, IAA was calculated to measure the quality of annotator training. At last, they made the corpus construction. The method of this study is a good demonstration of the construction a Chinese clinical corpus, however, similar to previous studies, it adopted the coarse-grained tagging patterns. Encouragingly, in 2019, Xiong et al[28] manually annotated a corpus about Chinese word segmentation (CWS) and Part-Of-Speech (POS) for Chinese clinical text at a fine-granularity level. It is undeniable that the work is an excellent reference, imperfectly, this study does not elaborate on the methods and implementary steps. In summary, building a corpus of Chinese clinical records has made some excellent beginnings.

### 2.2 TCM corpus construction

Compared with the corpus of western medicine in Chinese, the research on corpus construction of TCM clinical notes is too backward, and it is still in its infancy (*Table 1)*.. Fang et al[34] annotated the literature

corpus obtained from PubMed and developed an open database, TCMGeneDIT, to provide association information about TCMs, genes, diseases, TCM effects and TCM ingredients mined from vast amount of biomedical literature. This research method cannot be used in the study of clinical records. In respect of clinical records, Wang et al[16] constructed a type of annotated corpus for symptoms from chief complaint in TCM free texts. It is really an empirical study, but the text types, entity categories in the corpus were relatively few, and the authors did not list the detailed methods about guideline development and annotation work. Ruan et al[23] focused on symptoms and symptom-related entity extraction. Symptom was divided into TCM symptom and western symptom, and medicine was divided into TCM medicine and western medicine. The dataset was divided into two part, two experts annotated the symptom entity in the EMRs for training and testing a CRF model. Li et al[19] built a dataset of herb and symptom records, and annotated the relationship between symptoms and herbs. This study is meaningful for research on relation extraction from TCM health records. From the *Figure 1* we can know, there are large field of concepts in TCM clinical notes not involved, for example, the concept of meridians and collaterals, viscera, acupoints, etiology of TCM, syndrome elements, diagnosis methods, etc.

Taking the researches described above into consideration, the research on TCM clinical texts has some defects: first, there is no large corpus available in the TCM domain; second, only a small part of TCM concepts have been annotated systematically, while other types of entity have not been marked; third, the existing methods of TCM corpus construction are too rough, most of the previous studies have not described the implementation process of data selection, guideline drafting and annotation work. It is imperative to build a comprehensive corpus that follows a complete standard annotation scheme.

Figure 1 The top and second level of concepts in TCMLS.

# 3. Materials

## 3.1 Dataset

The dataset which contains 10197 records were from a modern Chinese case records database of TCM, it was transcripts of raw TCM clinical records collected by TCM doctors during their routine diagnosis and treatment work. The two main reason why we choose the transcripts data are introduced as follows:

(1) The transcript of TCM case records is an important resource for studying TCM. A complete case record of TCM contains abundant TCM knowledge, such as basic information, main complaint, syndrome differentiation, diagnosis, treatment or prescription, medicines, doses etc. Therefore, it has been an important resource of medical information for the study of unstructured documents, and the best document to obtain the academic thoughts and clinical experience of famous TCM experts[35–37]. Different from the RANs, case records of TCM are more refined, logical and enlightening[38]. Taking an example, the transcript *"□□,□□□,□□□□□,□□□□,□□□□□□□,□□,□□,□□□□,□□□,□□□",(cough, the phlegm of yellow thick is not easy to spit out, throat pain, traction pain in the position of EX-HN5 when cough, fear of cold, no sweat, fetid breath, yellow and greasy coating, slightly red tongue, not floating pulse), i*n which the key symptoms

for the diagnosis of TCM have been listed in this brief description, including four basic diagnosis procedures(inspection, listening and smelling, inquiry, and palpation)[39].

(2) It is of clinical significance to extract the knowledge hidden in massive TCM clinical texts and to become distilled in a concise form. A good example is the discovery of artemisinin which was spotted from the records of TCM, it is a medical advance that has saved millions of lives across the globe[40]. More recently, more and more studies have found that the diagnostic methods of TCM can help the disease diagnosis of modern medicine. For example, as for tongue diagnosis, it is found that tongue features identified in predicting early-stage breast cancer (BC)[41,42], geographic tongue is associated with disease severity and may be a marker of the psoriasis severity[43]. With regard to pulse diagnosis, Wang et al[44] found that there is significant difference between the pulse diagnosis signals of healthy volunteers and patients with fatty liver disease (FLD) and cirrhosis. String-like pulse in the left hand is always closely related to the liver disease[45].

In the light of the above reasons, we selected this transcripts data and aimed to establish a fine-grained entity corpus, with this corpus, more data mining technology can be applied to practice, so that more knowledge of clinical records of Chinese medicine will be excavated.

### 3.2 Entity selection

The method of selecting entities is rarely mentioned in previous studies. In our work, the process of entity selection is described in detail. Firstly, we analyzed the characteristics of our dataset. Secondly, 100 records are randomly selected for each annotator to establish the entity labels and annotate respectively. After this step, there were 26,10 and 46 concepts been marked by three annotators. Finally, three annotators discussed together about the inconsistent labels repeatedly and decide what entity type will be included in our study. After the above entity selection process, 4 annotators' cognition of four entity categories ("body part", "tongue diagnosis", "pulse diagnosis", "direction and position") is more consistent than others. In order to improve the work efficiency and quality, we chose the four entity categories rather than all types of TCM entities that occur in the dataset. There are some important concepts not involved in our study, for example, "symptom", "temporal word", "herbal medicine", etc. They will be the subjects in our following researches.

The four categories involving 13 entities in our experiment are of great significance to the pathogenesis analysis, syndrome differentiation, diagnosis and treatment, etc. Taking some examples, in the phrase "□□□□"(dispersing stagnated liver qi for promoting bile flow), "□"(liver) should be annotated as entity "Zang-organ", "□"(gallbladder) should be annotated as entity "Fu-organ". Here entity "□"(liver) and "□"(gallbladder) reflects the key Zang-Fu organs in the treatment procedure. In the Chinese word "□□□"(pain at the point of LI15), "□□"(LI15) should be annotated as entity "acupoint" which belongs to Large Intestine meridian(LI). Here "□□□"(pain in LI15) indicates the pathogenesis is the abnormality of meridian qi of LI, meanwhile, "□□"(LI15) is also a common point to be used to treat the shoulder pain. Moreover, tongue diagnosis and pulse diagnosis are indispensable information for the diagnosis of TCM, for instance, when a particular pulse appeared at the wrong place or in the wrong season, a serious disequilibrium of the system was

indicated[46], the tongue body mainly reflects the deficiency or excess of qi and blood of Zang-Fu organs, while the change of tongue coating is mainly used to judge the depth and severity of pathogenic qi[47]. For example, in the transcript "□□,□□,□□,□□□□,□□□□,□□□,□□,□□□,□□□" (insomnia, bitter taste, fond of drink, slightly yellow sputum, dry stool, hot breath, dryness of eyes, yellow and dry coating, thready and weak pulse), the yellow and dry coating reflects the disturbance of pathogenic heat internally, thready and weak pulse indicates the deficiency of healthy qi. Furthermore, the position and direction have important clinical diagnostic value. For example, according to TCM theory, different positions of the tongue correspond to five Zang organs, for example, the root of the tongue correspond to the kidney, top of the tongue correspond to the heart, the "□□□"(red tip of tongue) is probably one of the manifestations of heart fire.

3.3 Key and difficult points in entity annotation.

In our study, the fine-grained annotation methods are first used in the TCM clinical records. At the same time, the entities in our research, such as "acupoints", "Zang-Fu organs", "tongue manifestations" and "pulse conditions" etc. are not annotated in previous studies. Consequently, the annotation work is facing great challenges. The key and difficult points are introduced in the following part.

(1) Clinical narratives are often written in a medical sub-language with semantic categorization of words, domain specific terminology, incomplete phrases and omission of information[48]. TCM case records are written by practicing doctors, their brief forms appear very similar to ancient Chinese texts, and can only be understood by a professional doctor with the background of TCM. For example, in the transcript "□□,□□,□□" (short breath, stuffy nose especially in the left side), "□□"(especially in the left side) are brief form of "□□□□□□"( especially serious in the left side) which might be a complete form of expression. In the transcript "□□□,□□,□□"(thin and slippery coating, thready pulse especially in the left side), "□□"(especially obvious in the left side) are brief form of "□□□□□□□□"(the thready pulse is obvious in the left hand). In above two examples, "□□" has different meaning in different contexts in the transcripts.

(2) Fine-grained annotation is the foremost key and difficult part in our work, it is further divided the coarse-grained entity into finer subcategories. Kirk Roberts et al[49] found out that high-quality, fine-grained natural language annotations is important on the system's ability to recognize heart disease risk factors. Fine-grained annotated corpus will better support the NER tasks which automatically processing and analyzing EMRs. We collected the researches on entity tagging in recent years (*Table 1)*,, most of them are coarser-grained annotation, for example, "□□□" (left upper abdomen) was whole annotated as body part. However, in our fine-grained annotation guideline, most of words are shorter than two Chinese characters. For instance, "□□□" (left upper abdomen) have three entities, "□"(left), "□"(upper) and "□"(abdomen). Entity "□"(left) and entity "□"(upper) express the exact position of the abdomen, entity "□"(abdomen) is a part of the body. One major difference between Chinese and English text is that words in Chinese are formed by continuous Chinese characters without any space, and the boundary between fine-grained entities is not clear, as a result, fine-grained annotation on TCM clinical notes is time-consuming and requires an in-depth understanding of the document by the annotators.

# 4. Methods

4.1 Entity definition

In this study, we have summarized 4 types of data categories. 13 entity types are derived from the 4 categories. Referring to the concepts definitions of TCM in WHO International Standard terminologies on traditional medicine in the Western Pacific region[50], and the text book of diagnostics of Traditional Chinese Medicine[47], the definitions of annotated entities are listed in the *Table 2:*

Table 2 The definition and examples of the 13 entities in our study.

| Entity type | Definition | Examples（bold and underline） |
|---|---|---|
| **“Ordinary body part”** | the entity enables us to locate the exact positions of symptoms or medical tests, or the location of the disease. | □痒 (Itchy in the eyes), |
| **“Tongue body”** | the musculature and vascular tissue of the tongue, also referring to tongue substance. It is annotated only in the situation followed by the specific description of the tongue body manifestation. | □红，苔黄，脉滑 (red tongue, yellow coating, slippery pulse) |
| **“Tongue coating”** | a layer of moss-like material covering the tongue, also called tongue fur. It is annotated only in the situation if it is followed by the description of tongue coating manifestation. | 舌红，□黄，脉滑 (red tongue, yellow coating, slippery pulse) |
| **“Pulse”** | the radial artery of the wrist which include three sections: cun, guan, chi. The pulse entity is annotated only in the situation that it is followed by the description of the pulse condition. | 舌红，苔黄□滑 (red tongue, yellow coating, slippery pulse) |
| **“Acupoints”** | the point where a needle is inserted and manipulated in acupuncture therapy. | □□痛 (pain in LI15) |
| **“Meridians” and “collaterals”** | a system of conduits through which qi and blood circulate, connecting the bowels, viscera, extremities, superficial organs and tissues, making the body an organic whole, the same as channels and networks, also called meridians or channels, in short. | 左大腿□□□固定痛 (fixed pain in the stomach channel of foot-yangming of left leg) |
| **“Zang organs”** | an internal organ where essence and qi are formed and stored, including heart, liver, spleen, lung and kidney, also called five viscera. | 一直服调□化湿药 (always take the medicine of regulating spleen and removing dampness) |
| **“Fu organs”** | an internal organ where food is received, transported and digested, including gallbladder, stomach, large intestine, small intestine, urinary bladder and triple energizers, also called six bowels. | |
| **“Both the tongue body and tongue coating”** | the words of referring to tongue body as well as tongue coating. | □阿 (normal tongue) |

| | | |
|---|---|---|
| **“Tongue body manifestation”** | the specific description of tongue body manifestation, including tongue color, shape, sublingual vein etc. | □□□□□□□□ (red tongue, yellow coating, slippery pulse) |
| **“Tongue coating manifestation”** | the specific tongue of coating manifestation, including color, thickness, texture etc. | □□□□□□□□ (red tongue, yellow coating, slippery pulse) |
| **“Pulse condition”** | it is the specific description of arterial pulsation in TCM when pulse felt on examination. | □□□□□□□□ (red tongue, yellow coating, slippery pulse) |
| **“Direction and position”** | the description of direction and position which enable us to know the specific location of the body part. | □□□□□□ (pain of left knee joint) |

4.2 Annotation tools

In order to make the fine-grained marking process easier and more efficient, we have developed an entity annotation tool. As shown in *Figure 2,* the annotated Chinese characters were labeled with the predefined tags with a specific color. By setting the color of the label, we can distinguish the content of continuous annotations, and make the inconsistency more eye-catching. It will facilitate the modification of the annotations and the recording of the problems. Annotators are able to cancelled and re-annotated incorrect annotation in function column.

Figure 2 Details of the annotated page.

4.3 fine-grained annotation introduction

The principle of fine-gained annotation is splitting coarse entity to subcategories as much as possible. During the whole annotating process, what we use is the principle of fine-gained marking which can capture more context information. In our experiment, the Chinese words are further divided into the smallest semantic unit. For example, as shown in *Figure 3,* “□”(foot), “□”(knee) and “□”(limbs) should be annotated as “ordinary body part”, “□”(right) and “□”(below) should be separately annotated as “direction and position”. “□”(thin) should be annotated as “tongue coating manifestation”.

Figure 3 Examples of fine-grained annotation.

4.4 Annotation method

Taking previous research methods as references, based on the frequently-used process, we designed a refined and replicable method to developed the fine-grained annotation guideline and construct a fine-grained entity corpus by the following four steps (*Figure 4).*. During the whole process, the physicians play

the roles of guideline designers, annotators, and domain experts. The NLP researchers provide technical support and build annotation systems.

1. Determine the entity to be marked: this step has been elaborated in detail in the entity selection part.
2. Draft guideline development: after referred to some existing well-developed guidelines[21,51,52], a team of three annotators with the knowledge background of TCM randomly selected 100 records from the dataset to make the sample annotation. At the same time, they summarize the characteristics of included entities and drafted the annotation guideline. After repeated discussion, they drafted a fine-grained annotation guideline, in which different cases were exemplified for fine-grained entity annotation.
3. Guideline updating and consistency assessment: iterative annotation method was proposed to train the annotators and update the guideline. In each round, 100 unannotated records were randomly selected from the dataset. The guideline was constantly updated until the IAA met the standard of satisfaction (>0.9) which meant the markings of three annotators had achieved high consistency. If not, the iterative fine-grained annotation on the sample records would be continued. In the fourth round, the IAA () was greater than 0.9 finally, indicating that three annotators gradually come to a similar marking ability. During this step, detailed fine-grained annotation guideline was developed (in Appendix). Based on the draft guideline, we added more examples and supplemented with detailed explanations.
4. Corpus construction: on the basis of guideline developed in step 2 and step 3, three annotators commenced the annotation work independently. The dataset was divided into three parts, three annotators marked different part separately to reduce the consumption of time and improve annotation efficiency. During this period, we keep the annotation work as independent as possible, and the following annotation principles were formulated which will be strictly followed: (1) Although there are practical standards for medical records writing, sometimes errors exist in these texts. Wrongly written characters will not be annotated in any situations. For example, in the word “□□”(foot finger), “□”(finger) is miswritten and should be write as “□”(toes). So here “□”(finger) will not be annotated. (2) Entity annotation is allowed to be nested but not overlapped. For example, “□□□□□”(The body part where the fingers and palms are connected) should be annotated as “ordinary body part”, at the same time, “□”(finger) and “□”(palm) will be annotated as “ordinary body part” separately. (3) For some complex or ambiguous situations, we may make some appointments to unify the decisions specially. For example, there is a certain controversy that “□”(heart) in the word “□□”(palpitation) should be annotated as “ordinary body part” or “Zang organ”. “□□”(palpitation) is a subjective sensation of rapid and forceful beating of the heart, meanwhile it is a symptom name in TCM. It makes sense whether “□”(heart) was annotated as “ordinary body part” or “Zang organ”. After discussions, here “□”(heart) is unified to be annotated as “Zang organ”. (4) Punctuation should not be included in the annotation as far as possible. This is to minimize the interference of punctuations on the annotated entities. The punctuations will be annotated if it in the entity which cannot be separated anymore. For example, in word “□□□□”(slightly yellow coating on the tongue), “□”(slightly) is a description of the degree of yellow,

the formal expression should be “□□□”. Therefore in this case, “□, □” should be annotated as “tongue coating manifestation”.

In addition, during the comprehensive fine-grained annotation process, some measures were taken to ensure the quality: (1) Annotators are required to record the uncertain annotations, and discussed regularly until all the ambiguities achieving agreement. (2) Three annotators with the TCM knowledge background improved the marking accuracy and reduced the occurrence of uncertain cases. (3) Duplicate documents were assigned to three annotator groups in corpus construction step for the IAA evaluation in order to measure the quality of all annotations.

Figure 4 Workflow for guideline development and fine-grained entity corpus construction, IAA, inter-annotator agreement.

4.5 Inter-annotator agreement

The calculation of IAA(often known outside of corpus linguistics as inter-rater agreement) is motivated by the need to deal with the problem of subjectivity in judgments about things that are not observable with the senses[53]. Cohen’s kappa is the coefficient of internal consistency which is a widely used index for assessing IAA. It is appropriate for nominal and ordinal data, where there are two or more raters per subject. In our study, we choose the Cohen’s Kappa statistic to measure the consistency of three annotators’ work.

$$\kappa = \frac{P_0 + P_e}{1 - P_e}$$

$P_0$ is the observed agreement between two annotators, $P_e$ is the chance agreement between the annotators if each annotator randomly picked a category for each annotation. $P_e$ is computed from a contingency matrix representing agreements and disagreements. The calculation method of IAA refers to Tang et al [54] and Carletta [55]. According to the method, we calculate the consistency between two of the three annotators. The consistency is considered to be satisfactory when all the three value are greater than 0.9 at the same time.

# 5. Results And Discussion

5.1 Annotation consistency

We added duplicates (600 records) in each annotator’s tagging task for the purpose of calculating the IAA. The result showed that the IAA value during corpus construction step remained at a relatively high level (0.93, 0.94, 0.94) (last column of *Figure 5)*.. The IAA evaluation shows that this fine-grained entity corpus is of good quality.

As shown in *Figure 5,* our marking task was a repeated time-consuming work, in which the whole marking process was carried out for 5 rounds. In the fourth round, IAA values reached more than 0.9, indicating that the three annotators had a high degree of consistency in the understanding of labels and TCM records, and they had ability to accomplish these annotation tasks with satisfactory consistency. Furthermore, the IAA value in each annotation round is higher than that of the previous round which show that our method of iterative annotations and discussions is effective. As a group, the approaches and detailed implementation steps in our study is an excellent reference of the future research.

Figure 5 The IAA () in the first to four round annotation training and final corpus construction between two of the three annotators (W, Y, Z).

Table 3 The entity counts and annotation counts and its corresponding percentage.

| Entity classification | Entity type | Total entity counts | Total annotation counts | Percentage in the corresponding type (entity/annotation) |
|---|---|---|---|---|
| Body part | Ordinary body part | 462 | 21093 | 75.3% / 56.3% |
| | Pulse | 22 | 6148 | 3.6% / 16.4% |
| | Tongue coating | 10 | 4978 | 1.6% / 13.3% |
| | Tongue body | 7 | 3789 | 1.1% / 10.1% |
| | Acupoints | 87 | 469 | 14.2% / 1.3% |
| | Zang-organ | 5 | 139 | 0.8% / 0.4% |
| | Meridians and collaterals | 16 | 34 | 0.98% / 0.1% |
| | Fu-organ | 2 | 3 | 0.3% / 0.008% |
| | Both tongue body and coating | 2 | 793 | 0.3% / 2.1% |
| | Total | 613 | 37446 | 100% / 100% |
| Tongue manifestation | Tongue coating manifestation | 102 | 10911 | 38.9% / 72.7% |
| | Tongue body manifestation | 160 | 4088 | 61.1% / 27.2% |
| | Total | 262 | 14999 | 100% / 100% |
| Pulse condition | Pulse condition | 90 | 9573 | 100% / 100% |
| Direction and position | Direction and position | 139 | 5781 | 100% / 100% |
| Total counts | 13 | 1104 | 67799 | |

Table 4 Examples of top 10 entities in each entity type.

| Entity type | Total counts | Entity examples (top 10) |
|---|---|---|
| Ordinary body part | 21091 | □(mouth)(2252), 头(head)(1853), 腹(abdomen)(1689), 胃(stomach)(1267), 喉(larynx)(962), 腰(893)(waist),肢(limbs)(686), 背(back)(585), 身(body)(583), 手(hand)(578) |
| Pulse | 6148 | 脉(pulse)(6091), 尺脉(chi pulse)(11), 肾脉(kidney pulse)(10), 关(guan)(6), 寸(cun)(6), 尺(chi)(4), 关脉(guan pulse)(4), 肝(liver)(2), 沉取(taking heavily)(1), 脉沉取(taking the pulse heavily)(1) |
| Tongue coating | 4978 | 苔(coating)(4765), 舌苔(tongue coating)(188), 舌(tongue)(16) |
| Tongue body | 3789 | 舌(tongue)(3695), 舌质(tongue body)(87), 苔(tongue coating)(3), 舌苔(tongue coating)(1), 舌头(tongue)(1), 质(tongue body)(1) |
| Acupoints | 469 | 风池(GB20)(66), 太阳穴(EX-HN5)(51), 肩井(GB21)(40), 大椎(DU14)(30), 环跳(GB30)(27), 肩髃(LI15)(14), 少海(HT3)(12), 委中(BL40)(11), 承扶(BL36)(11), 天宗(SI11)(10) |
| Zang-organ | 139 | 心(heart)(125), 肺(lung)(5), 肾(kidney)(4), 脾(spleen)(3) |
| Meridians and collaterals | 34 | 膀胱经(Bladder meridian, BL)(8), 胃经(Stomach meridian, ST)(6), 大肠经(Large intestine meridian, LI)(4), 肝经(Liver meridian, LI)(2), 足太阳(Bladder meridian, BL)(2), 心经(Heart meridian, HT)(1), 肺经(Lung meridian, LU)(1),手阳明经(Large intestine meridian, LI)(1), 足少阳(Gallbladder meridian, GB)(1), 小肠经(Small intestine meridian, SI)(1) |
| Fu-organ | 3 | 胆(gallbladder)(2), 胃(stomach)(1) |
| Both tongue body and coating | 793 | 舌(tongue)(793) |
| Tongue coating manifestation | 10911 | 薄(thin)(3612), 黄(yellow)(1907), 腻(slimy)(1725), 干(791)(dry), 白(738)(white), 略黄(slightly yellow)(570), 少(365)(less), 厚(254)(thick), 润(233)(moist), 滑(150)(slippery) |
| Tongue body manifestation | 4088 | 红(red)(893), 淡(pale)(564), 略红(slightly red)(467), 暗(dark)(216), 略暗(slightly dark)(216), 红暗(red and dark)(195), 齿印(teeth-marked)(144), 暗红(dark and red)(127), 淡暗(pale and dark)(126), 略淡(slightly pale)(122) |
| Pulse condition | 9573 | 细(thready)(3493), 弦(string-like)(1364), 弱(faint)(841), 沉(sunken)(651), 滑 (slippery)(616), 数(534)(rapid), 软(soft)(473), 平(normal)(420), 略弦(slightly string-like)(180), 略数(slightly rapid)(123) |
| Direction and position | 5781 | 左(left)(1262), 右(right)(1110), 下(lower)(736), 上(upper)(282), 心(center)(273), 中(middle)(199), 尖(tip)(193), 前(front)(141), 外(outside)(136), 外侧(outward)(128) |

Table 5 Examples of top 10 annotated acupoints in corresponding meridians.

| Meridians | Annotated acupoints | Examples |
| --- | --- | --- |
| Lung meridian (LU) | 鱼际(LU10)(3), 云门(LU2)(3) | “右□痛”(pain in right LU2) |
| Large intestine meridian (LI) | 肩髃(LI15)(14), 曲池(LI11)(9), 合谷(LI4)(8), 臂臑(LI14)(7), 手三里(LI10)(5), 巨骨(LI16)(4), 肘髎(LI12) (1) | “左□消失”(Left LI15 pain disappears) |
| Stomach meridian (ST) | 解溪(ST41)(5), 脾关(ST31)(4), 气冲(ST30)(2), 梁丘(ST34)(2), 下关(ST7)(2), 内庭(ST44)(1), 足三里(ST36)(1), 丰隆(ST40)(1), 人迎(ST9)(1) | “左膝,□,坐起时痛甚”(pain in ST41 in the left knee were aggravated when sitting up) |
| Spleen meridian (SP) | 血海(SP10)(1), 三阴交(SP6)(1), 大横(SP15)(1), 腹结(SP14)(1) | “下肢麻痹,左,□以下,膝痛”(paralysis of left lower extremities, below the SP10, gonyalgia) |
| Heart meridian (HT) | 少海(HT3)(12) | “左锁骨头痛,右□痛”(pain in left collarbone and right HT3) |
| Small intestine meridian (SI) | 天宗(SI11)(10), 秉风(SI12)(5), 曲垣(SI13)(2), 肩贞(SI9)(1), 秉风穴(SI12)(1), 天容(SI17)(1) | “右□痛,右前臂蚁行感”(right SI11 ache, a sense of ant movement in the right forearm) |
| Bladder meridian (BL) | 委中(BL40)(11), 承扶(BL36)(11), 白环俞(BL30)(10), 大肠俞(BL)(9), 承山(BL57)(7), 秩边(BL54)(3), 昆仑(BL60)(2), 通天(BL7)(2), 申脉(BL62)(1), | “右□、□痛,静脉阻塞”(ache in right BL40 and BL57 aside, vein occlusion) |
| Kidney meridian (KI) | 涌泉(KI1)(2), 太溪(KI3)(2), 然谷(KI2)(1) | “脚底热感,□”( a heat sensation in sole of the foot, KI1) |
| Pericardium meridian (PC) | 大陵(PC7)(1) | “痛点,左阳池,左少海,右□”(ache point, in left SJ4, HT3 and right PC7) |
| Triple energizer meridian (TE) | 阳池(TE4)(8), 耳门(TE21) (2), 肩髎(TE14)(2) | 手关节痛,□处,怕冷(hand joints ache, in the position of TE4, afraid of cold) |
| Gallbladder Meridian (GB) | 风池(GB20)(66),肩井(GB21)(40), 环跳 (GB30)(27), 居髎(GB29)(10), 阳陵泉(GB34)(5), 侠溪(GB40)(1), 维道(GB28)(1) | “现痛点,□下面”(at present, the ache point is in the GB20 below) |
| Liver meridian (LR) | 急脉(LR12)(2), 太冲(LR3)(2) | “便秘,右□痛”(constipation, pain in right LR12) |

| | | |
|---|---|---|
| Extra point (EX) | 太阳穴(EX-HN5)(51),太阳(EX-HN5)(6), 外膝眼(EX-LE5)(4), 夹脊(EX-B2)(3), 膝眼(EX-B6)(2), 鹤顶(EX-LE2)(1), 颈百劳(EX-UX8)(1), 腰眼(EX-B6)(1) | “头痛,□□□”(headache, in the EX-HN5) |
| Governor vessel (GV) | 大椎(GV14)(30), 腰阳关(GV3)(8), 长强(GV1)(1), 风府(GV16)(1), 前顶(GV21)(1) | “项背强痛,右□□垣甚”(stiff pain in the nape and back, especially in the right GV14 and SI13) |
| Conception vessel (CV) | 中脘(CV12)(1), 曲骨(CV2)(1) | “□□胁胁痛,吐白粘痰”(pain in the CV12 and costal region, spitting white and sticky phlegm) |

5.2 Annotation results and analysis

The fine-grained annotated corpus has 1104 entities and 67799 tokens totally. The distribution of entities and tokens are shown in the *Table 3.* The proportion of entity “ordinary body part”, “tongue body”, “tongue coating”, “tongue body manifestation”, “tongue coating manifestation”, “pulse”, “pulse condition” and “direction and position” are much higher than other entities. In “body part” category, the entity “ordinary body part” (21093) is the most, followed by entity “pulse” (6148), “tongue coating” (4978) and “tongue body” (3789). Among the entity “ordinary body part”, we noticed that some annotated entities are concepts of western medicine. For instance, “毛细血管充血状” (capillary vessel) and “椎间盘”(intervertebral disk) are body part concepts of western medical anatomy. Apparently, modern TCM clinical records has been greatly influenced by western medicine, and the modern case records of TCM are a combination of TCM and Western medicine knowledge. Then, the entities about “tongue body manifestation” (4088), “tongue coating manifestation” (10911) and “pulse condition” (9573) are relatively more. After reading the original text, we observed that almost every TCM case record documented the pulse or tongue diagnosis information. It can be seen that tongue diagnosis and pulse diagnosis are one of the most common diagnostic methods in TCM, in which the “tongue coating manifestation” (10911) has high diagnostic value in practice.

Moreover, the expressions of many concepts of TCM are not uniform, and there is a lot of entities that are similar in semantics but different in names, such as “腹”, “腹部” both means abdomen, “足阳明经” and “胃经” both means Stomach meridian (ST), “内” and “内侧” both means inside, “中心” and “中间” both refer to the center of a position. For such synonyms of different expressions, on the one hand, it would reduce the reliability of statistical analysis results of the corpus, on the other hand, this kind of expressions are real and raw Chinese language, it will increase the adaptability of machine learning models.

Table 6 The top 20 syndromes in the dataset.

| syndrome | count | syndrome | count |
|---|---|---|---|
| 血瘀(blood stasis) | 2773 | 气虚(qi deficiency) | 315 |
| 血虚(blood deficiency) | 1525 | 肝胃不和(liver-stomach disharmony) | 292 |
| 气滞(stagnation of the circulation of qi) | 1011 | 胃气滞(stomach qi stagnation) | 243 |
| 内热(inner heat) | 920 | 阴虚(yin deficiency) | 231 |
| 夹风(with wind evil) | 877 | 肝热(liver heat) | 213 |
| 肝郁(stagnation of liver-QI) | 689 | 夹热(accompany with hot) | 206 |
| 夹湿(accompany with dampness) | 571 | 肾虚(kidney deficiency) | 186 |
| 脾虚(spleen deficiency) | 552 | 气滞血瘀(qi-stagnancy and blood stasis) | 183 |
| 湿热(dampness-heat) | 409 | 肺气不宣(obstruction of pulmonary qi) | 155 |
| 风热郁肺(wind-heat stagnate the lung) | 375 | 阳虚(yang deficiency) | 139 |

*Figure 7* Examples of some relationships between top 10 syndromes and top 10 pulse conditions. The solid lines suggest that there are great possible relations between syndromes and pulse conditions. No line doesn't mean no relation between them.

*Figure 8* Examples of some relationships between top 10 syndromes and top 10 tongue body and coating manifestations. The solid lines suggest that there are great possible relations between them. No line doesn't mean no relation between them.

5.3 The top 10 syndromes and its relationship with of the entities of pulse and tongue body(coating) manifestations.

The top 20 syndromes, which had been preprocessed, are listed in *Table 6.* As an important part of TCM diagnosis, syndrome differentiation is through comprehensive analysis of symptoms and signs, which has implications for determining the cause, nature and location of the illness and the patient's physical condition[50]. By referring to the textbook of "diagnostics of Traditional Chinese Medicine"[47], solid lines are used to connect the probably related entities to syndromes, as shown in the *Figure 7* and *Figure 8.* From the *Figure7* we can see, there are many-to-one and one-to-many relations between syndromes and pulse conditions. For examples, string-like pulse is probably caused by qi stagnation and liver depression, the blood deficiency manifests as thready or faint pulse. In *Figure 8,* the blood stasis syndrome performs as the multiple clinical tongue body manifestations (dark, dark and red, red and dark), blood deficiency manifests as the pale tongue, the yellow coating may be the result of inner heat or dampness heat. As can be see, the pulse or tongue body(coating) manifestation of high frequency shows a close relationship with the syndrome of high frequency. Taking an instance, in the transcripts *"□□□□,□□,□□□,□□□"(feel tired from time to time, pale tongue body, thin and yellow coating, faint pulse in left hand),* and *"□□□□,□□□,□□,□□,□□,□□□,□□□,□□"(delayed menorrhea, fear of cold, fatigue, drowsy, pale tongue body, thin and moist coating, thready pulse),* after comprehensive analyzing the clinical manifestations, both syndrome of these two cases should be concluded as blood deficiency, the pale tongue body, faint and thready pulse are two of the important indications of blood deficiency syndrome.

There are also some exceptions. For example, blood stasis is the most frequent syndrome in our dataset. In TCM basic theory, blood stasis syndrome probably manifests as rough pulse, slow pulse, and tight pulse. [47] But interestingly, these three are not mentioned in the top 10 pulse conditions. To find out the cause of the result, we looked up the original text and noticed that in the TCM clinical free text, the patients of blood stasis syndrome may not appear as the above-mentioned pulse conditions. For instance, in the transcript *"□□□□□□□,□□□□□,□□,□□□□,□□□□,□□□,□□□,□□□,□□(postoperative cerebral vascular aneurysm, clouded in mind, dry lips, dark purple, sticky sputum, dry stool, clear urination, yellow and greasy coating, dark red tongue, slow pulse),* the syndrome of this case should be summarized as blood stasis with phlegm-heat, the postoperative cerebral vascular aneurysm and dark red tongue body reflects the stagnated blood inside the body, but the moderate pulse is not a typical pulse of blood stasis syndrome.

It can be seen that the main content of the corpus corresponds to the annotation results mostly, constructing a corpus helps to obtain and analyze the content of dataset. However, there are some cases which do not conform to this. That is because TCM is experience-based clinical medicine, its clinical cases are comprehensive and changeable. Although tongue diagnosis and pulse diagnosis have a certain diagnostic effect, only comprehensive analysis by the four examination methods can help the practitioner to diagnosis and treatment accurately.

5.4 Examples of special entities and analysis

In our fine-grained entity corpus, there are some special annotations need to be explained. In most cases, general rule is that there is a modification of direction and location word in front of the body part, such as “右下肢”(right lower limb), “左膝关节”(left knee joint). However, there are still some particular expression of TCM, for example, "少腹”(lower abdomen), "小腹”(lower abdomen) are two of the “ordinary body part” entities in TCM. In our annotation work, “少” and “小” should not be annotated as “direction and position” separately. In order to preserve the particular expression of TCM, the entities similar to the above two cases will not be split as well.

There are some entities of combinability attribute in TCM. Taking an example, the record "背心怕冷"(the center of the back are afraid of cold), in which “心” means the center position on the back rather than the heart viscera, so “心” should be annotated as “direction and position”, the record “心虚胆怯” (timidity due to insufficiency of qi and deficiency of blood of the heart), in which “心” should be annotated as Zang organ, moreover, “心” in the word”心肌”(cardiac muscle) refers to the anatomical heart, and should be annotated as “original body part”. In above three cases, the word “心” should be annotated as different entity type accordingly.

Furthermore, the case records of TCM has many abbreviations and polysemy, for example, the transcript “左尺尤”(especially in left chi), in which “尺”(chi) is the brief form of “尺脉”(chi pulse), here the word “脉”(pulse) was an ellipsis. Another example, the word “舌” ought to be annotated as “tongue body”(e.g. “舌红”(red tongue)), “tongue coating”(e.g. "舌腻”(tongue coating is slimy)), and “both tongue body and tongue coating” (e.g. “舌可”(normal tongue)) according to different contexts.

Some special entities are annotated as “direction and position” entity type, such as the records “□□□□”(two cun downward) and “□□4□□”(4 cun sideward). Cun is a common ancient unit of length (≈3.33 centimeter) especially in locating the acupoints or meridians. It is quite similar to ancient Chinese medical texts.

Hence the annotation of TCM clinical records complicated. It is quite different from the annotation work of western medical records before. Abundant relevant TCM knowledge is necessary for the annotators to analyze the meaning of the context.

5.5 Examples and analysis of entity type with less entity counts or annotation counts.

The entity types with low percentage of entity counts in TCM clinical records contain “pulse” (3.6%), “tongue coating” (1.6%), “tongue body” (1.1%), “Zang organs” (0.8%), “Fu organs” (0.3%), “meridians and collaterals” (0.98%), and “both tongue body and coating” (0.3 %), with low percentage of annotation counts include “acupoints” (1.3%), “Zang organs” (0.4%), “meridians and collaterals” (0.1%), “Fu organs” (0.008%), and “both tongue body and coating” (2.1 %). The results would be attributed to the possible reasons as follows:

(1) It is easy to form inertial thinking when annotating the entity “Body part” which resulting the entity “Fu-organs” is rare in the annotation result. For example, “□”(stomach) is annotated as “Fu-organ” for 2 times but as “ordinary body part” for 1267 times. The entity “□” is definitely annotated as “ordinary body part” in "□□□□□□□"(after subtotal gastrectomy), but should be potentially annotated as “Fu-organ” in “□□,□□□□,□□□□□□”(bitter taste, epigastric distension, regurgitation sometimes in evening). In-deepth analysis is required for annotators to distinguish whether “□”(stomach) should been annotated as "ordinary body part” or “Fu-organ”. From the perspective of TCM’s cognition of body structure, “□□” is a term of TCM, the word “□”(stomach) should be annotated as “Fu-organ” in “□□”(stomach), but from western medical perspective, here it should be annotated as “ordinary body part”. In our annotated corpus, “□”(stomach) was annotated as “ordinary body part” in most records. It is thus clear that the three annotators with TCM background are much influence by western medicine knowledge.

(2) The dataset in our study is Chinese medicine physician case records instead of acupuncture case records. Thus, the entity “acupoints” (1.3%), and “meridians and collaterals" (0.1%) account for very small proportion in our corpus. *Table 5* lists the examples of top 10 annotated acupoints and corresponding meridians. From *Table 5* we can see, there are more acupoints belonging to the meridians BL, GB, GV and EX, and the acupoints GB20(66), EX-HN5(51), GB21(40), GV14(30) are the most annotated entities. Interestingly, acupoints are mostly used to describe a symptom, especially symptom of pain. We can reasonably infer that different focus of knowledge field and clinical habits of TCM physicians may also lead to this result.

Furthermore, from *Table 3* we can see, there are not many entities about “tongue body” (7), “tongue coating” (10), “pulse” (22) and “both tongue body and tongue coating” (2), while they have large quantity of annotations (3789, 4978, 6148, 793). Hence one can see that the expressions of “tongue body”, “tongue coating”, “pulse” and “both tongue body and tongue coating” are relatively consistent and frequently-used in TCM clinical records.

# 6. Conclusions And Future Work

The corpus construction is a fundamental and indispensable task for groping for NLP technics to automatic recognition of TCM valuable knowledge. In this paper, we have successfully presented a method of building a fine-grained annotated entity corpus based on case record of TCM. This method introduced detailed steps as well as the implementing process, involving data selection, draft guideline development, iterative annotations for guideline updating, consistence assessment, and corpus construction. High IAA value was achieved finally in our annotation work, indicating that our approaches are effective and the corpus is of high quality. This work lays a solid foundation for future TCM corpus construction and NER researches.

There are still some inevitable shortcomings in our work, such as the entity types were not comprehensive enough. Because of the limitation of time, we can't complete the marking of all existent entities in our dataset. In the future, we will annotate more entity types, such as symptoms and prescriptions, to enrich the guideline and corpus by using the method introduced in this paper. More types of TCM clinical record from different sources will also be annotated to improve the applicability of the corpus. Furthermore, based on the corpus, we will develop more corresponding algorithms to support the NLP techniques. Last point, deep research of the polysemy, abbreviation, relationship between entities are also the next focuses in our further research work.

# Abbreviations

IAA: inter-annotator agreement; TCM: Traditional Chinese Medicine; NER: Named Entity Recognition; EMR: Electronic Medical Record; MOH: Ministry of Health; NLP: Natural language processing; TE: BI-RADS: Temporal expression; Breast Imaging Reporting and Data System; i2b2: Integrating Biology and the Bedside; RANs: Resident Admit Notes; CWS: Chinese word segmentation; POS: Part-Of-Speech; BC: breast cancer; FLD: fatty liver disease; LU: Lung meridian; LI: Large intestine meridian; ST: Stomach meridian; SP: Spleen meridian; HT: Heart meridian; SI: Small intestine meridian; BL: Bladder meridian; KI: Kidney meridian; PC: Pericardium meridian; TE: Triple energizer meridian; GB: Gallbladder Meridian; LR: Liver meridian; EX: Extra point; GV: Governor vessel; CV: Conception vessel.

# Declarations

Ethics approval and consent to participate.

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (grant number 61801058, 61501063).

Authors' contributions

WYQ and YY guided the whole research work, WXF developed the annotation tool and calculated annotation consistency, YY, ZTT, YYF were responsible for the annotation schema development and the whole annotation work. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Dr Jiang for providing the Chinese case records of TCM.

# References

[1] Basic Specification for Eelectronic Medical Records (Trial). China's health quality management 2010; 17: 22–23.

[2] Jane Qiu. Traditional medicine: a culture in the balance. Nature 2007; 448: 126.

[3] Nadkarni P. M，Ohno-Machado L and Chapman W. W. Natural language processing: an introduction. J Am Med Inform Assoc 2011; 18: 544–51.

[4] Lei J，Tang B，Lu X，Gao K，Jiang M and Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc 2014; 21: 808–14.

[5] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 2004; 32: 267–70.

[6] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in Health Technology & Informatics 2006; 121: 279.

[7] Vivian Law，Craig Knox，Yannick Djoumbou，Tim Jewison，Chi Guo An，Yifeng Liu，Adam Maciejewski，David Arndt，Michael Wilson and Vanessa Neveu. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Research 2014; 42: 1091–7.

[8] Xu Y，Wang Y，Liu T，Liu J，Fan Y，Qian Y，Tsujii J and Chang E. I. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. J Am Med Inform Assoc 2014; 21: e84–92.

[9] He B，Dong B，Guan Y，Yang J，Jiang Z，Yu Q，Cheng J and Qu C. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. J Biomed Inform 2017; 69: 203–217.

[10] Gao Y，Gu L，Wang Y，Wang Y and Yang F. Constructing a Chinese electronic medical record corpus for named entity recognition on resident admit notes. BMC Med Inform Decis Mak 2019; 19: 56.

[11] Cai X，Dong S and Hu J. A deep learning model incorporating part of speech and self-matching attention for named entity recognition of Chinese electronic medical records. BMC Med Inform Decis Mak 2019; 19: 65.

[12] Liu Z，Tang B，Wang X，Chen Q，Li H，Bu J，Jiang J，Deng Q and Zhu S. CMedTEX: A Rule-based Temporal Expression Extraction and Normalization System for Chinese Clinical Notes. AMIA Annu Symp Proc 2016; 2016: 818–826.

[13] Chapman Wendy W, Nadkarni, Prakash M, Hirschman, Lynette, D\”Avolio, Leonard W, Savova, Guergana K, Uzuner, Ozlem. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association 2011; 540–543.

[14] Yang JF，Yu QB ，Guan Y and Jiang ZP. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. Acta Automatica Sinica 2014; 40: 1537–1562.

[15] Wang YQ，Yu ZH ，Jiang YG ，Liu YC ，Chen Li and Liu YG. A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records. Journal of Biomedical Informatics 2012; 45: 210–223.

[16] Wang Y，Yu Z，Chen L，Chen Y，Liu Y，Hu X and Jiang Y. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study. J Biomed Inform 2014; 47: 91–104.

[17] Wang H，Zhang W，Zeng Q，Li Z，Feng K and Liu L. Extracting important information from Chinese Operation Notes with natural language processing methods. J Biomed Inform 2014; 48: 130–6.

[18] Cao C，Sun M and Wang S. Extracting terms from clinical records of traditional Chinese medicine. Front Med 2014; 8: 347–51.

[19] Li Y. B，Zhou X. Z，Zhang R. S，Wang Y. H，Peng Y，Hu J. Q，Xie Q，Xue Y. X，Xu L. L，Liu X. F and Liu B. Y. Detection of herb-symptom associations from traditional chinese medicine clinical data. Evid Based Complement Alternat Med 2015; 2015: 270450.

[20] Xu D，Zhang M，Zhao T，Ge C，Gao W，Wei J and Zhu K. Q. Data-Driven Information Extraction from Chinese Electronic Medical Records. PLoS One 2015; 10: e0136270.

[21] Zhang S.，Kang T.，Zhang X.，Wen D.，Elhadad N. and Lei J. Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models. J Biomed Inform 2016; 60: 334–41.

[22] Wan H，Moens M. F，Luyten W，Zhou X，Mei Q，Liu L and Tang J. Extracting relations from traditional Chinese medicine literature via heterogeneous entity networks. J Am Med Inform Assoc 2016; 23: 356–65.

[23] Ruan T，Wang M，Sun J，Wang T，Zeng L，Yin Y and Gao J. An automatic approach for constructing a knowledge base of symptoms in Chinese. J Biomed Semantics 2017; 8: 33.

[24] Zhang Y and Wang X. Clinical Named Entity Recognition From Chinese Electronic Health Records via Machine Learning Methods. JMIR Med Inform 2018; 6: e50.

[25] Miao S，Xu T，Wu Y，Xie H，Wang J，Jing S，Zhang Y，Zhang X，Yang Y，Zhang X，Shan T，Wang L，Xu H，Wang S and Liu Y. Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. JMIR Med Inform 2018; 119: 17–21.

[26] Bao X. Y，Huang W. J，Zhang K，Jin M，Li Y and Niu C. Z. A customized method for information extraction from unstructured text data in the electronic medical records. Beijing Da Xue Xue Bao Yi Xue Ban 2018; 50: 256–263.

[27] Wang Q，Zhou Y，Ruan T，Gao D，Xia Y and He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. J Biomed Inform 2019; 92: 103133.

[28] Xiong Y，Wang Z，Jiang D，Wang X，Chen Q，Xu H，Yan J and Tang B. A fine-grained Chinese word segmentation and part-of-speech tagging corpus for clinical text. BMC Med Inform Decis Mak 2019; 19: 66.

[29] Chapman W. W，Nadkarni P. M，Hirschman L，D'Avolio L. W，Savova G. K and Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association 2011; 18: 540–543.

[30] Kim J. D，Ohta T，Tateisi Y and Tsujii J. GENIA corpus—semantically annotated corpus for bio-textmining. Bioinformatics 2003; 19 Suppl 1: i180.

[31] Karimi Sarvnaz，Metke-Jimenez Alejandro，Kemp Madonna and Wang Chen. C adec: A corpus of adverse drug event annotations. Journal of Biomedical Informatics 2015; 55: 73–81.

[32] Proceedings of the 2012 Workshop on Biomedical Natural Language Processing[C]. 2012:

[33] Herrero-Zazo María，Segura-Bedmar Isabel，Martínez Paloma and Declerck Thierry. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of Biomedical Informatics 2013; 46: 914–920.

[34] Fang Y. C□Huang H. C□Chen H. H and Juan H. F. TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. BMC Complement Altern Med 2008; 8: 58.

[35] Wang FX and Li J. Studying the Medical record of Traditional Chinese Medicine is the best way for School inheritors to acquire the academic experience of famous Teachers Journal of Pediatrics of Traditional Chinese Medicine 2019; 15: 8–11.

[36] Li Zhenji. □He Xingdong.□Wang Sicheng. and Xu Chunbo. Strategic Thought on Clinical Experience and Academic Thoughts of Famous Old Chinese Medicine Doctors. World Chin Med 2012; 7: 1–4.

[37] Meng Qingyun. On the Value, Characteristics and Research Methods of the Traditional Chinese Medicine. Journal of Traditional Chinese Medicine 2006; 568–570.

[38] Z Xiaoping. Traditional Chinese Medical Record Science Bei Jing: China Press of Traditional Chinese Medicine, 1995.

[39] Gao Z and Dong JC. From four TCM diagnostic methods used in combination to precision TCM syndrome-based treatment. China Journal of Traditional Chinese Medicine and Pharmacy 2019;

[40] Miller L. H and Su X. Artemisinin: discovery from the Chinese herbal garden. Cell 2011; 146: 855–8.

[41] Lo L. C□Cheng T. L□Chiang J. Y and Damdinsuren N. Breast cancer index: a perspective on tongue diagnosis in traditional chinese medicine. J Tradit Complement Med 2013; 3: 194–203.

[42] Lo L. C□Cheng T. L□Chen Y. J□Natsagdorj S and Chiang J. Y. TCM tongue diagnosis index of early-stage breast cancer. Complement Ther Med 2015; 23: 705–13.

[43] Picciani B. L□Souza T. T□Santos Vde C□Domingos T. A□Carneiro S□Avelleira J. C□Azulay D. R□Pinto J. M and Dias E. P. Geographic tongue and fissured tongue in 348 patients with psoriasis: correlation with disease severity. Scientific World Journal 2015; 2015: 564326.

[44] Nanyue W□Youhua Y□Dawei H□Bin X□Jia L□Tongda L□Liyuan X□Zengyu S□Yanping C and Jia W. Pulse Diagnosis Signals Analysis of Fatty Liver Disease and Cirrhosis Patients by Using Machine Learning. Scientific World Journal 2015; 2015: 859192.

[45] Wang Ya and Fan Xiaoxuan. Study on the Relationship between Left String-like Pulse and Liver Disease. Journal of Emergency in Traditional Chinese Medicine 2015; 24: 1193–1194.

[46] Bedford D. E. The Ancient Art of Feeling the Pulse. Br Heart J 1951; 13: 423–37.

[47] Zhu WF Diagnostics of Traditional Chinese Medicine China Press of Traditional Chinese Medicine, 2007.

[48] Raghavan P Fosler-Lussier E and Lai A.M. Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise. AMIA Annu Symp Proc 2012; 2012: 1366–74.

[49] Roberts K Shooshan S. E Rodriguez L Abhyankar S Kilicoglu H and Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform 2015; 58 Suppl: S111–9.

[50] WHO International Standard terminologies on traditional medicine in the Western Pacific region, 2010 http://www.wpro.who.int/publications/who_istrm_file.pdf. Accessed 1 Oct 2019.

[51] Uzuner O South B. R Shen S and DuVall S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011; 18: 552–6.

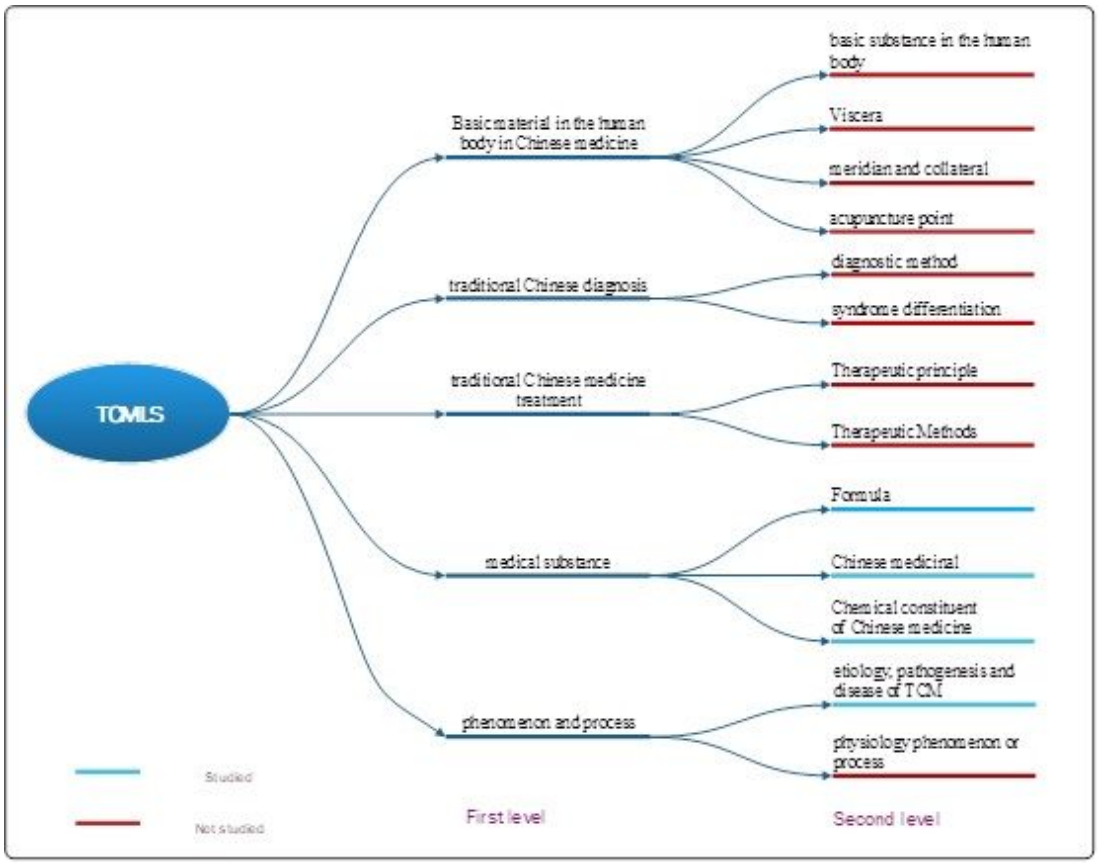
[52] Fan J. W Yang E. W Jiang M Prasad R Loomis R. M Zisook D. S Denny J. C Xu H and Huang Y. Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. J Am Med Inform Assoc 2013; 20: 1168–77.

[53] Boguslav M and Cohen K. B. Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing. Stud Health Technol Inform 2017; 245: 298–302.

[54] Tang W Hu J Zhang H Wu P and He H. Kappa coefficient: a popular measure of rater agreement. Shanghai Arch Psychiatry 2015; 27: 62–7.

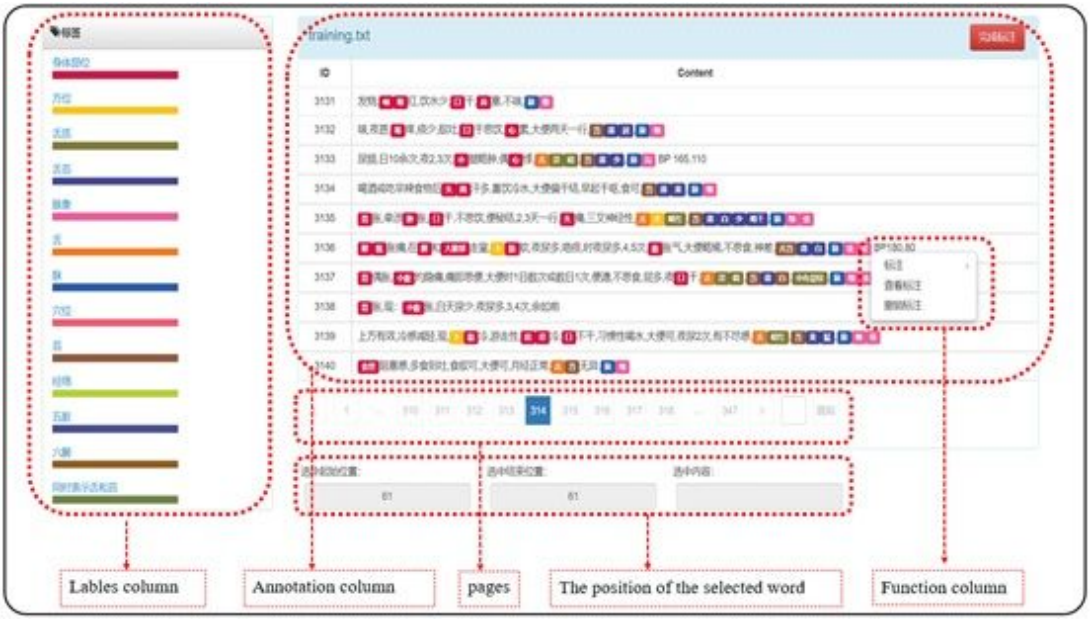
[55] Carletta Jean. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 1996; 22: págs. 249–254.

# Figures



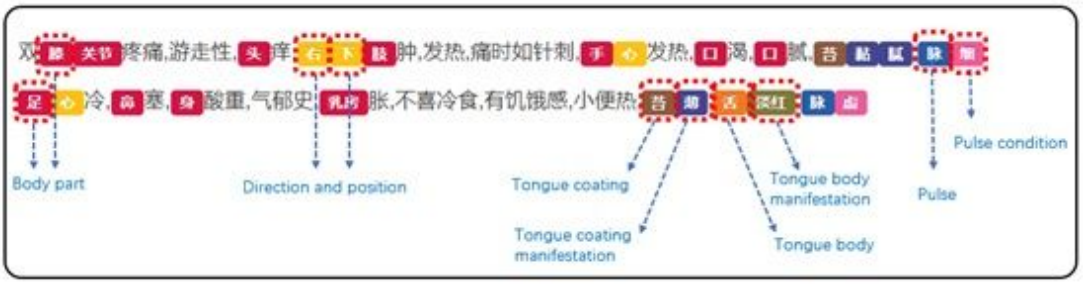

## Figure 1

The top and second level of concepts in TCMLS.



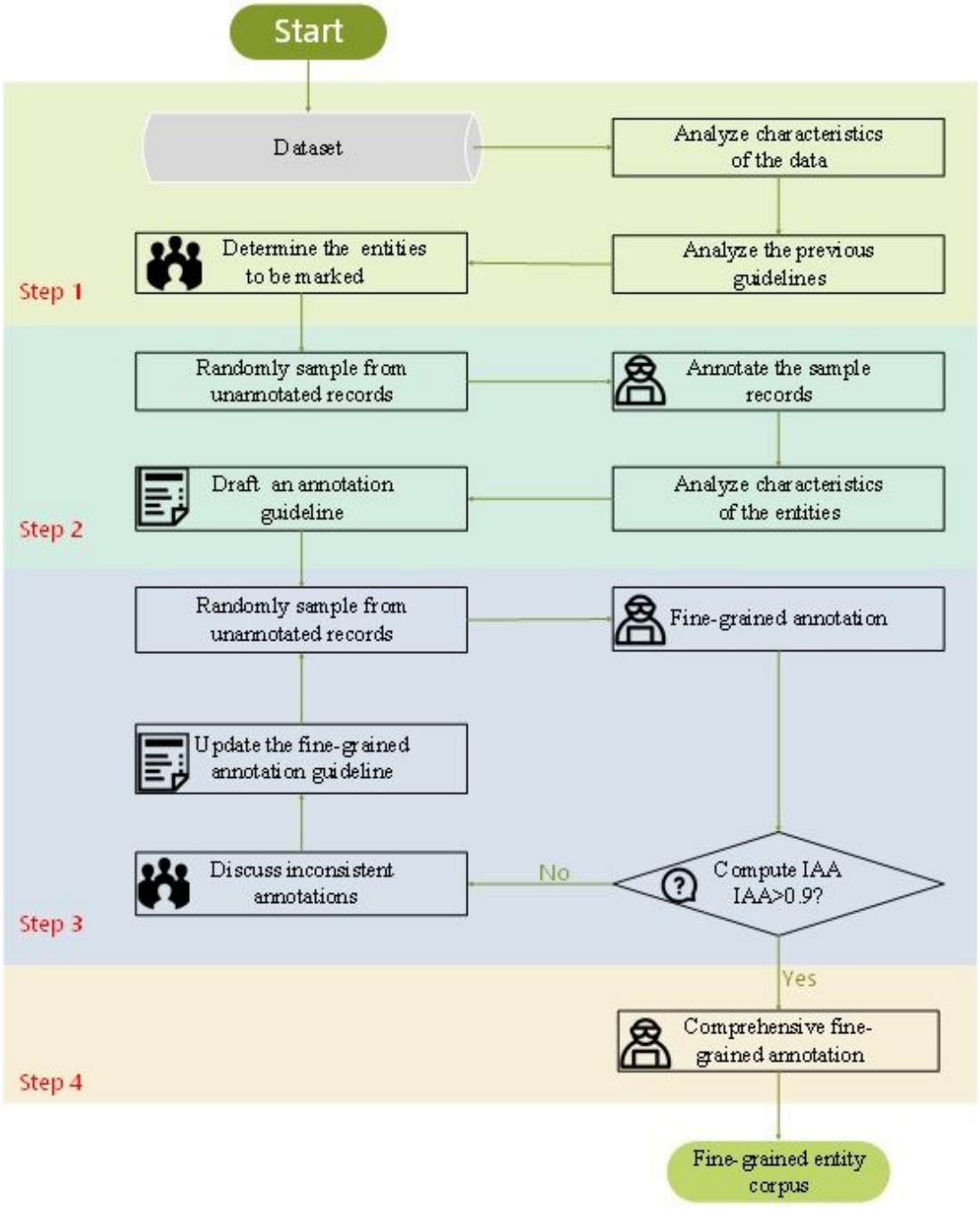

## Figure 2

Details of the annotated page.




## Figure 3

Examples of fine-grained annotation.




## Figure 4

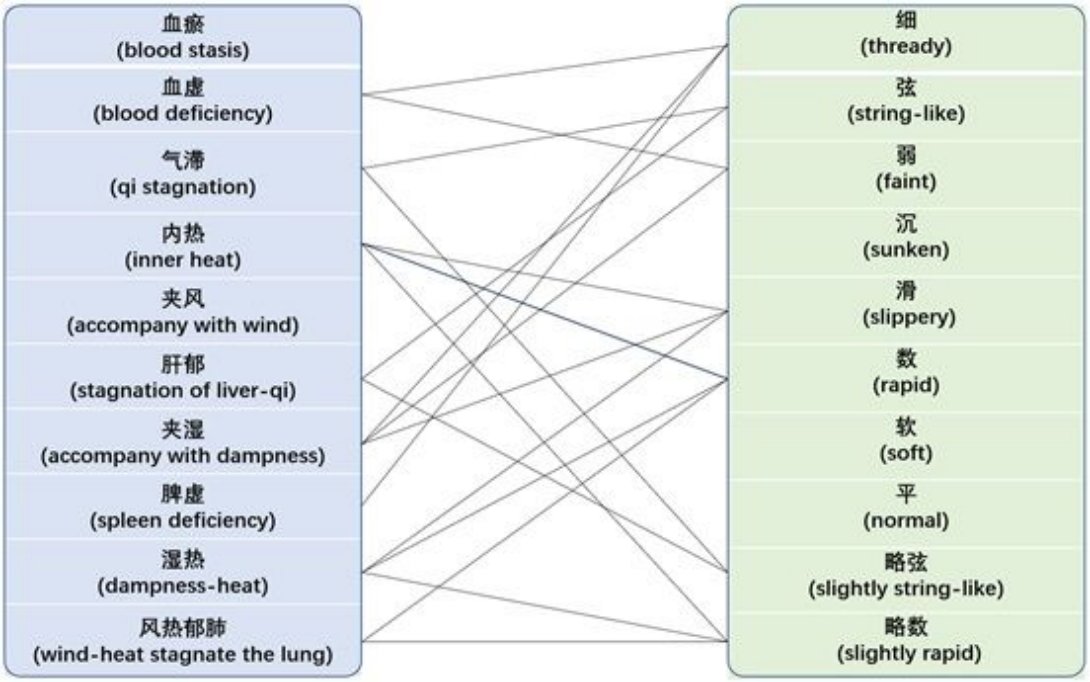
Workflow for guideline development and fine-grained entity corpus construction, IAA, inter-annotator agreement.


**Figure 5**

The IAA (□) in the first to four round annotation training and final corpus construction between two of the three annotators (W, Y, Z).

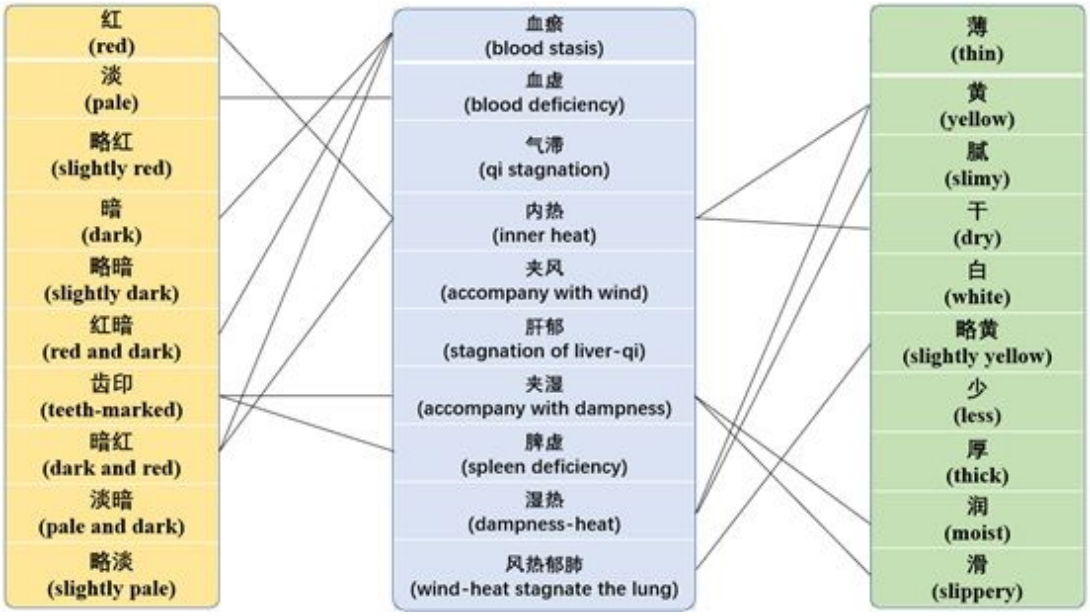
**No Figure 6 was included in this version.**

**Figure 6**




## Figure 7

Examples of some relationships between top 10 syndromes and top 10 pulse conditions. The solid lines suggest that there are great possible relations between syndromes and pulse conditions. No line doesn't mean no relation between them.




## Figure 8

Examples of some relationships between top 10 syndromes and top 10 tongue body and coating manifestations. The solid lines suggest that there are great possible relations between them. No line doesn't mean no relation between them.