Original Research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–17
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241286260
journals.sagepub.com/home/dhj
Sage

Construction, evaluation, and application of an electronic medical record corpus for cerebral palsy rehabilitation

Meirong Xiao[1,2], Qiaofang Pang[2], Yean Zhu[1], Lang Shuai[1] and Guoqiang Jin[1]

Abstract

Objective: The electronic medical records (EMRs) corpus for cerebral palsy rehabilitation and its application in downstream tasks, such as named entity recognition (NER), requires further revision and testing to enhance its effectiveness and reliability.

Methods: We have devised an annotation principle and have developed an EMRs corpus for cerebral palsy rehabilitation. The introduction of test-retest reliability was employed for the first time to ensure consistency of each annotator. Additionally, we established a baseline NER model using the proposed EMRs corpus. The NER model leveraged Chinese clinical BERT and adversarial training as the embedding layer, and incorporated multi-head attention mechanism and rotary position embedding in the encoder layer. For multi-label decoding, we employed the span matrix of global pointer along with softmax and cross-entropy.

Results: The corpus consisted of 1405 EMRs, containing a total of 127,523 entities across six different entity types, with 24,424 unique entities after de-duplication. The inter-annotator agreement of two annotators was 97.57%, the intra-annotator agreement of each annotator exceeded 98%. Our proposed baseline NER model demonstrates impressive performance, achieving a F1-score of 93.59% for flat entities and 90.15% for nested entities in this corpus.

Conclusions: We believe that the proposed annotation principle, corpus, and baseline model are highly effective and hold great potential as tools for cerebral palsy rehabilitation scenarios.

Keywords

Electronic medical record, cerebral palsy, named entity recognition, medical entity corpus, information extraction

Submission date: 11 October 2023; Acceptance date: 3 September 2024

Introduction

Cerebral palsy (CP) is a group of central motor and postural developmental disorders caused by nonprogressive interference in the developing brain. The prevalence of CP is approximately 3.4‰ in low- and middle-income countries and 1.6‰ in high-income countriess.[1] CP prevalence among individuals aged 0–18 in China was found to be 2.07‰, with a higher prevalence in rural areas compared to urban areas (2.75‰ vs. 1.90‰). From 2008 to 2019,

[1]Department of Rehabilitation Medicine, First Affiliated Hospital of Nanchang University, Nanchang, China
[2]Bioengineering College, Chongqing University, Chongqing, China

Corresponding authors:
Lang Shuai, Department of Rehabilitation Medicine, First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi 330006, China.
Email: shuailangncc@126.com

Guoqiang Jin, Department of Rehabilitation Medicine, First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi 330006, China.
Email: gq_jin18@outlook.com

the prevalence showed a mean annual increase of 38.13%,[2] establishing CP as a disorder with a significant disease burden among Chinese children.[3] Addressing motor dysfunction should be the primary focus of clinical rehabilitation therapy and research,[4] as the main symptoms of CP involve abnormal motor development and posture. The knowledge base for motor function rehabilitation in CP can reduce the clinical stress experienced by physicians. It can also improve the capacity of health services and the standard of living for children with CP in low- and middle-income countries, especially in rural areas. The electronic medical records (EMRs) of patients with CP encompass abundant diagnostic and treatment information pertinent to motor function and a wealth of insights provided by physicians. Therefore EMRs are the main data source for building the knowledge base. Nevertheless, the unstructured free text within EMRs necessitates extracting structured information, such as constructing corpus and named entity recognition (NER) task, for subsequent analysis and utilization. Despite the crucial role they play in advancing knowledge for motor function rehabilitation in CP, the construction of an EMRs entity corpus and the task of NER are often overlooked and are not typically implemented.

The existing medical entity corpus presents considerable opportunities for further enhancement and refinement. Firstly, the scarcity of large-scale medical entity corpus is exacerbated by the high cost of annotation and challenges in obtaining EMRs, leading to most available corpus being limited in scale[5] or sourced from official websites[6] and literature[7] rather than real EMRs. Subsequently, the granularity of entity delineation in several Chinese EMRs corpus does not yet meet practical needs, such as the CCKS2017 Task 2 and CCKS2018 Task 1,[4,8] in which the symptom entity *right lower limb shortening deformity* is simply divided into body entity *right lower limb* and symptom entity *shortening deformity*. Given the significant emphasis on body structure and motor function levels in CP, motor function rehabilitation is necessary.[9] Using nested entities to represent medical concepts across various levels of granularity enables physicians to focus more precisely on body structures, the aforementioned example needs to be labeled as the nested body entity *right lower limb* within the symptom entity *right lower limb shortening deformity*. Furthermore, modified entities can facilitate doctors in assessing the motor function level of CP, such as *severely reduced muscle tone* or *slightly reduced muscle tone*. However, the majority of existing corpus of Chinese EMRs[8,10,11] have not incorporated modified entities. Thirdly, many studies verify only the consistency among annotators,[5,12] that is, inter-annotator agreement (IAA (2,2)), but they often overlook the consistency of the annotators' own annotations. This oversight may compromise the quality and credibility of the corpus. Therefore, such a corpus, with a suitable scale and high credibility based on real EMRs, would greatly facilitate clinical applications related to CP and provide valuable insights for research and development in the field.

NER is a natural language processing technology that extracts clinical concepts, including diseases, symptoms, drugs, and other medical named entities, from EMRs. Several machine learning-based algorithms, such as hidden Markov model (HMM)[13] and conditional random field (CRF),[14] have been applied to NER. Compared to machine learning-based approaches, deep-learning-based methods can automatically extract features, thereby minimizing the need for feature engineering. Therefore, the current dominant paradigm for NER is the deep-learning-based approach, such as the bidirectional long short-term memory (BiLSTM)[15] and BiLSTM-CRF model.[16] Further, within deep-learning models, pre-trained models based on transformer[17] are initially trained and unsupervised on extensive text data to acquire generic word representations, which are subsequently fine-tuned using domain-specific data. This approach can notably enhance the model's efficacy in low-resource tasks and facilitate adaptation to downstream applications.[18] The bidirectional encoder representation from transformers (BERT)[19] model is the most prominent pre-trained model in the academic community. Its variants include bertcner in the clinical area of Chinese, as well as the BERT-base in the general domain of Chinese.[20] When recognizing medical entities, the clinical BERT performs better than the general BERT.[21] The BERT-BiLSTM-CRF framework, which integrates the BERT model with BiLSTM-CRF, currently stands as the most commonly employed framework in NER task.[20] Other researchers have also employed multi-head attention mechanism to obtain better results.[8,12] In contrast to the CRF, the global pointer (GP) network has the capability to recognize nested entities.[22] Given the ability of the attention mechanism to dynamically adjust attention weights for capturing critical information and the proficiency of the GP in parsing nested entities, it is reasonable to assume that a NER model based on Chinese clinical BERT, the attention mechanism and GP network can achieve excellent performance in recognizing nested entities within the EMRs of the CP.

Therefore, to address the existing lack of an EMRs corpus for CP, we created a high-quality corpus with an improved scale and suitable for clinical tasks. We developed the principles of named entity annotation for EMRs of CP under the direction of medical experts and controlled the quality of the annotations throughout. The EMRs corpus of CP was developed using 1405 EMRs from 281 patients with CP. Moreover, we compared the performance of HMM, CRF, BiLSTM, BERT, attention mechanism, and other approaches to discover a NER model suitable for this corpus. The experiment proves that the Chinese clinical BERT + multi-head attention mechanism + GP strategy outperformed other strategies and was effective in recognizing nested medical entities.

Methods

Corpus construction

The procedure for constructing the EMRs entity corpus is illustrated in Figure 1, including the establishment of the annotation principles, pre-annotation, and formal annotation.

Annotation principles. Creating reasonable annotation principles and strictly adhering to them during the annotation procedure are the most crucial steps in constructing the EMRs entity corpus. In Step 1 (shown in Figure 1), with the assistance of medical experts, the initial draft of the annotation principles was created with reference to the 2010 i2b2 annotation guidelines,[23] Chinese medical text annotation guidelines,[24] and the Unified Medical Language System (UMLS) semantic types. Exploring the recognition of modified entities is essential because modified information is crucial for clinicians to assess the severity of a patient's condition. Therefore, six types—SYMPTOM, BODY, TREATMENT, DISEASE, CHECK, and ADJUNCT—were used in the annotation principles to classify the medical entities. Further details on entity boundaries and types, as well as general annotation principles, can be found in Appendix A.2.

Preparatory work. Data preparation: The flow chart for inclusion and exclusion of patients with CP is shown in Figure 2. The inclusion criteria were as follows: patients must meet all three of the following criteria: (a) having a discharge diagnosis of CP, (b) receiving inpatient treatment, and (c) CP listed as the first-listed diagnosis. The exclusion criteria were as follows: patients are excluded if they meet one of the following criteria: (a) five categories of EMRs are incomplete, including *medical history characteristics*, *discharge summary*, *history of present illness*, *diagnosis basis*, and *hospital course*, or (b) the presence of other motor function diseases that could significantly affect motor function. As our study utilizes historical EMRs and does not involve the collection of blood or other biological samples, written informed consent from each patient was not required. Instead, the study was reviewed and approved by the ethics committee, which waived the need for individual consent given the nature of the data used. After obtaining approval from the ethics committee of the Children's

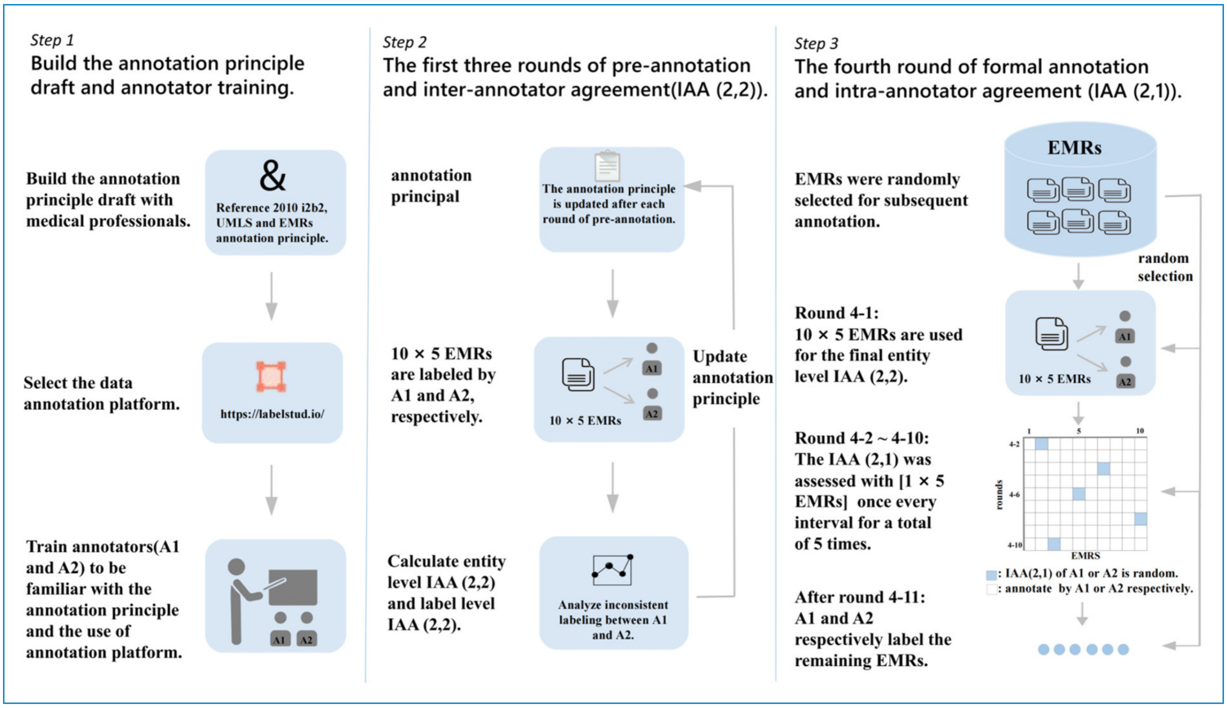

Step 1
Build the annotation principle draft and annotator training.
Build the annotation principle draft with medical professionals.
&
Reference 2010 i2b2, UMLS and EMRs annotation principle.
Select the data annotation platform.
https://labelstud.io/
Train annotators(A1 and A2) to be familiar with the annotation principle and the use of annotation platform.
A1
A2
Step 2
The first three rounds of pre-annotation and inter-annotator agreement(IAA (2,2)).
annotation principal
The annotation principle is updated after each round of pre-annotation.
10 × 5 EMRs are labeled by A1 and A2, respectively.
10 × 5 EMRs
A1
A2
Update annotation principle
Calculate entity level IAA (2,2) and label level IAA (2,2).
Analyze inconsistent labeling between A1 and A2.
Step 3
The fourth round of formal annotation and intra-annotator agreement (IAA (2,1)).
EMRs were randomly selected for subsequent annotation.
EMRs
random selection
Round 4-1:
10 × 5 EMRs are used for the final entity level IAA (2,2).
10 × 5 EMRs
A1
A2
Round 4-2 ~ 4-10:
The IAA (2,1) was assessed with [1 × 5 EMRs] once every interval for a total of 5 times.
1
5
10
4-2
4-6
4-10
rounds
EMRS
🟦: IAA(2,1) of A1 or A2 is random.
□: annotate by A1 or A2 respectively.
After round 4-11:
A1 and A2 respectively label the remaining EMRs.

Figure 1. Overall flow chart of corpus construction. (Step 1) Under the guidance of medical experts and by referencing existing annotation guidelines, an initial draft of entity annotation principles for CP EMRs was developed. After the annotators (A1 and A2) became familiar with the annotation principles, a portion of EMRs were chosen for training on the annotation platform. (Step 2) Three rounds of pre-annotation were then conducted. In each round of pre-annotation, two annotators independently labeled the same 10 × 5 EMRs. The annotation principles were refined based on evaluations of inter-annotator agreement (IAA (2,2)) and analyses of discrepancies between the annotators. (Step 3) The objective of the fourth round is to focus on intra-annotator agreement (IAA (2,1)) and the annotation of the remaining EMRs. First, 10 × 5 EMRs were selected to assess the final IAA (2,2). Then, [1 × 5 EMRs] were randomly selected to each annotator, two annotators repeated the labeling of the [1 × 5 EMRs] five times without the annotator's awareness. The IAA (2,1) was calculated across these five annotations to evaluate the consistency of each annotator's entities over time.

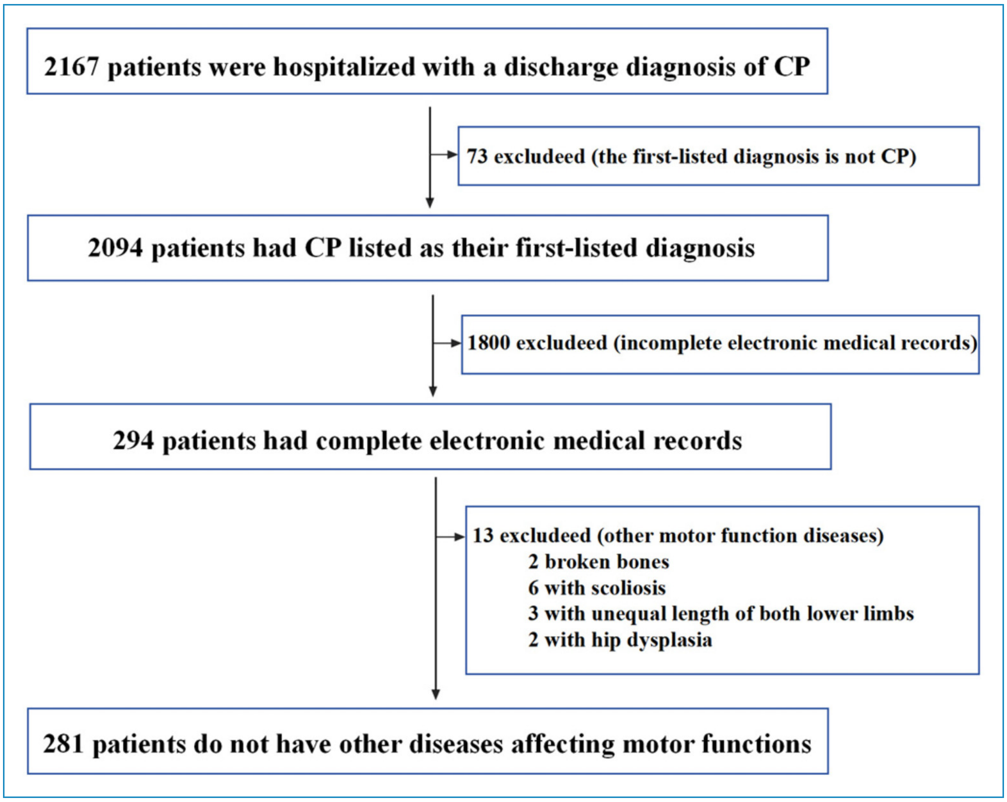

2167 patients were hospitalized with a discharge diagnosis of CP
73 excludeed (the first-listed diagnosis is not CP)
2094 patients had CP listed as their first-listed diagnosis
1800 excludeed (incomplete electronic medical records)
294 patients had complete electronic medical records
13 excludeed (other motor function diseases)
2 broken bones
6 with scoliosis
3 with unequal length of both lower limbs
2 with hip dysplasia
281 patients do not have other diseases affecting motor functions

Figure 2. Flowchart for selection criteria of patients with cerebral palsy (CP).

Hospital Attached to Chongqing Medical University, 281 patients who satisfied both the inclusion and exclusion criteria were selected for this study. Of the 281 patients, 100 patients were hospitalized more than once, resulting in additional EMRs. However, given that the EMRs of the same patient across different times are relatively similar, and to balance the cost of labeling with the diversity of the corpus, this study included only the EMRs from the first hospitalization of each patient. With each patient contributing five categories of EMRs, so we have 1405 EMRs (281 × 5 EMRs, 281 represents the number of patients and 5 represents five categories EMRs of each patient). Prior to labeling, sensitive data from the EMRs, such as the patient's name, ID number, contact information, residential address, and doctor's name, were removed through data desensitization.

Personnel training: In this study, two annotators (marked as A1 and A2), both master's degree candidates in medical fields, were involved in the annotation process. After the annotators (A1 and A2) became familiar with the annotation principles, a portion of EMRs were chosen for labeling training on the annotation platform (https://labelstud.io/). The labeling training aimed to familiarize annotators with the annotation principles and the proper usage of annotation platform.

Pre-annotation. The quality of a corpus can be evaluated in terms of its consistency and size. In contrast to other corpus, we implemented a more rigorous approach to consistency control. We conducted IAA (2,2) to assess agreement between annotators, and we also verified intra-annotator agreement (IAA (2,1)) for each individual annotator. By incorporating these measures, we aimed to establish standardized corpus. We initially developed a draft of the annotation principles specific to EMRs oriented towards CP rehabilitation. Through multiple iterations of annotation and revision, we ensured a satisfactory level of consistency in the annotated data.

As shown in Figure 1, a pre-annotation procedure is executed to ensure a high level of IAA (2,2) in Step 2. At this stage, we performed random sampling without replacement on a subset of 30 × 5 EMRs out of the total 281 × 5 EMRs. These sampled EMRs were evenly divided into three groups to complete three rounds of pre-annotation. The pre-annotation procedure was as follows:

1. Two annotators independently labeled the same 10 × 5 EMRs.
2. The IAA (2,2) was assessed, and the inconsistencies between the two annotators were analyzed.
3. Under the guidance of medical experts, the annotation principles were updated.
4. The process then returns to step (1) to initiate the subsequent round of pre-annotation based on the most updated annotation principles.

As the annotation principles in the pre-annotation phase were still being updated, the 30 × 5 EMRs were relabeled in the formal annotation phase. In this study, 10 × 5 EMRs from each round proved adequate for achieving satisfactory IAA (2,2); however, future studies should ascertain the appropriate pre-annotation size on a case-by-case basis.

Formal annotation. In step 2, the annotation principles tended to be stable after three rounds of pre-annotation. Subsequently, the formal annotation phase was launched. To further determine whether the annotation principles were stable, we last assessed the IAA (2,2) during the round 4-1 of formal annotations. Specifically, 10 × 5 EMRs were drawn from the remaining 251 × 5 EMRs by sampling without replacement. Two annotators independently annotated the same 10 × 5 EMRs, and the IAA (2,2) was calculated based on the annotation results.

In Step 3 (Figure 1), we focused on the IAA(2, 1) of each annotator. The concept of test-retest reliability[25] was used to assess IAA (2,1). 172 × 5 EMRs were randomly drawn from the remaining 241 × 5 EMRs using sampling without replacement, and subsequently these 172 × 5 EMRs were equally allocated to A1 and A2 for annotation in round 4-2 to 4-10. Each annotator will completed a total of 86 × 5 EMRs for annotation. The mathematical formula of this annotation process was 86 × 5 EMRs = (10 × 5 EMRs) * 4 + (9 × 5 EMRs) * 5 + [1 × 5 EMRs] * 5. *N indicates the number of rounds of annotation that need to be completed, whereas *4 signifies that four rounds of annotations need to be completed. The [1 × 5 EMRs] were randomly selected to each annotator, two annotators repeated the labeling of the [1 × 5 EMRs] five times without the annotator's awareness. To avoid retrieval practice effect, the order of [1 × 5 EMRs] in which each round was randomized. IAA (2,1) was represented by the consistency of the annotators' results for the [1 × 5 EMRs] across the five rounds of annotations.

Finally, the remaining 69 × 5 EMRs and 30 × 5 EMRs from the pre-annotation phase were formally annotated by two annotators. The labeling process took place at the Children's Hospital of Chongqing Medical University and spanned a total of 6 months.

NER methods

The deep-learning-based NER model typically consists of an embedding layer, a contextual encoder, and a tag decoder.[18] The proposed baseline model utilized the Chinese clinical BERT + adversarial training (AT) as the embedding layer, while the encoder layer incorporated multi-head attention mechanism and rotary position embedding (RoPE). For multi-label decoding, the span matrix of GP, along with softmax and cross-entropy, were employed. The network structure of the model is illustrated in Figure 3.

Embedding layer: Since the introduction of BERT,[19] pre-trained models for downstream tasks have gained popularity because they can learn universal language representations. In our model, we utilized a Chinese clinical BERT called bertcner[20] as the base embedding. During the AT process, small perturbations were introduced to the training samples, allowing the neural network model to adapt and increase its robustness by learning to handle these variations. Therefore, we added perturbations in the gradient direction specifically to the embedding layer of the Chinese clinical BERT. Suppose that the embedding vector of a Chinese clinical BERT is $X = [x_1,\ x_2,\ \cdots,\ x_n]$, perturbation ($\Delta x$) is added to original X to obtain the new vector X_{adv}, and is used as the final embedding vector.

$$X_{\text{adv}} = X + \Delta x \tag{1}$$

Context encoder: In text processing tasks, multi-head attention mechanism can extract additional text features from multiple perspectives and levels. In our model, the number of multi-head attentions corresponding to the number of entity types. The particular entity type a in the input text of length n is encoded and transformed to obtain the sequence vectors $Q = [q_{1,a}, q_{2,a},\ \cdots, q_{n,a}]$ and $K = [k_{1,a}, k_{2,a},\ \cdots, k_{n,a}]$. The inner product of the vectors produces the attention score (equation (2), which serves as the scoring function for segments i to j corresponding to entities of type a).

$$s_a(i, j) = q_{i,a} \top k_{j,a} \tag{2}$$

RoPE is utilized to incorporate positional information in the attention mechanism, leveraging both boundary information and contextual relationships between words.[26] RoPE uses the nature of complex number operations to achieve relative position embedding in the form of absolute position embedding. Specifically, it constructs matrices and corresponding to positions i and j, respectively, satisfying the relation $R_i \top R_j = R_{j-i}$, thereby obtaining a new score (equation (3)). Meanwhile, RoPE exhibits strong extrapolation capability and is capable of effectively handling sequences of random lengths,[26] which is beneficial for the subsequent processing of long texts.

$$s_a(i, j) = (R_i q_{i,a}) \top (R_j k_{j,a}) = q_{i,a} \top R_i \top R_j k_{j,a} = q_{i,a} \top R_{j-i} k_{j,a} \tag{3}$$

Tag decoder: In our model, the most widely used CRF decoder is replaced by the span matrix of GP.[27] Decoding the output of the attention mechanism module with a span matrix of GP allows for the decoding of nested entities. For a sentence of length n, the tag decoder architecture creates z span matrices of size $n \times n$, z indicates the number of entity types. The row index of mark 1 is the head position of the entity, and the column index of mark 1 is the tail position; therefore, the head and tail boundaries of the entity can

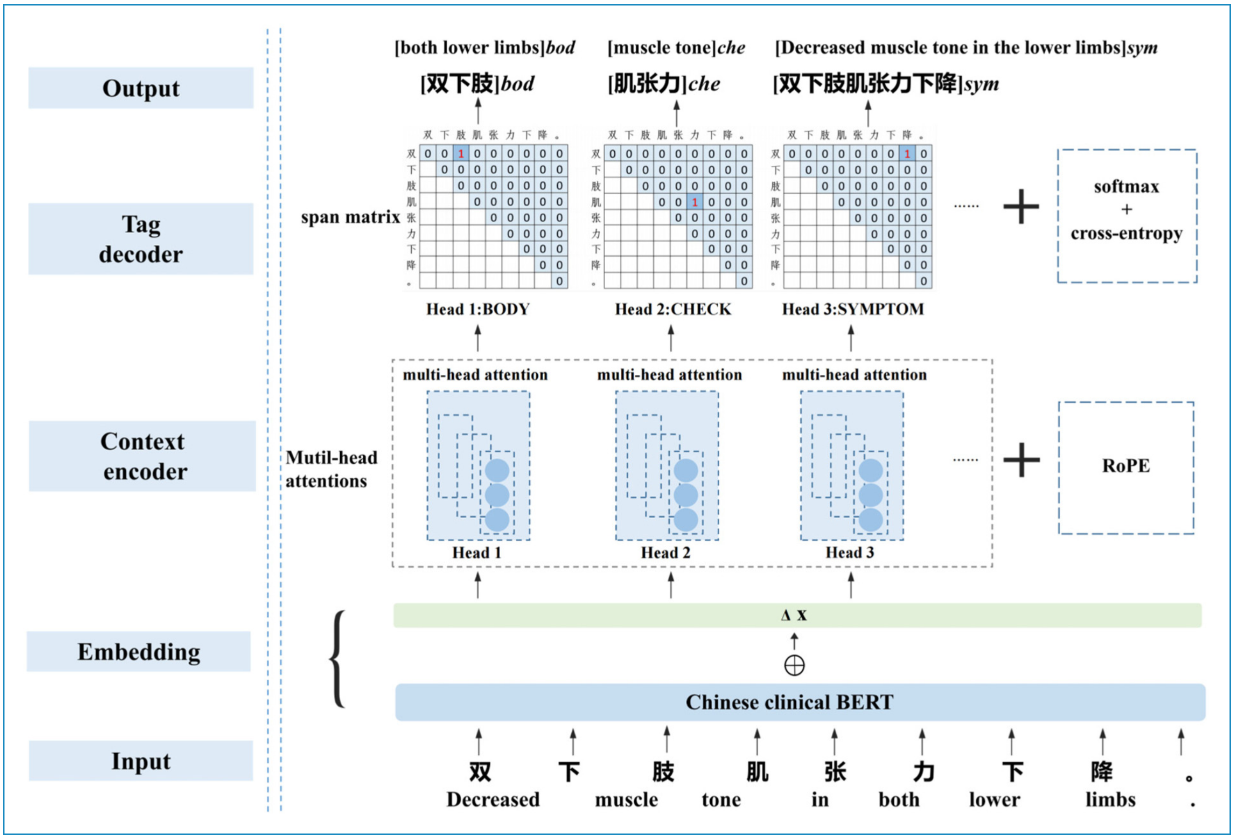


Figure 3. Network structure of the NER model. The leftmost section of the figure is labeled with the names of the model modules, while on the right side, the corresponding details for an example are provided. Take the character string “双下肢肌张力下降” (Decreased muscle tone in the lower limbs.)” as input, the model outputs the result of entity recognition: [双下肢]*bod*([both lower limbs]*bod*), [肌张力]*che*([muscle tone]*che*), and [双下肢肌张力下降]*sym* ([Decreased muscle tone in the lower limbs]*sym*).

be decoded directly by mark 1 simultaneously. In the absence of restrictions on entity length and the allowance for nested entities, a sentence of length n has at most $n(n+1)/2$ candidate entities, and f is the real number of given entities in the sentence, indicating that model need to select these f target entities from $n(n+1)/2$ candidate entities. The simplest approach is to use a sigmoid activation function, which transforms into a $n(n+1)/2$ binary classification problem. However, $n(n+1)/2 \gg f$, which leads the problem of severe category imbalance. The single-label classification problem with softmax and cross-entropy does not suffer from category imbalance. To solve the category imbalance problem, Su et al.[28] proposed an extension of softmax and cross-entropy to multi-label classification problems. This method reduces the multi-label classification problem to the difference between the target and non-target scores. The loss function for identifying the entities of type a is given by equation (4).

$$\log\left(1+\sum_{(i,j)\in P_a} e^{-s_a(i,j)}\right)+\log\left(1+\sum_{(i,j)\in Q_a} e^{s_a(i,j)}\right) \qquad (4)$$

where P_a is the set of all entities of type a (target) of this sample and Q_a is the set of all non-entities or not type a entities (non-target) of this sample. Note that we only need to consider combinations with $i \le j$.

Results

Results of corpus construction

Consistency evaluation of corpus. The consistency was calculated using the F-score. The annotation result of one annotator (A1) was considered the standard annotation, and the precision (P), recall (R), and F-score of the annotation result of another annotator (A2) were calculated. The consistency was calculated using equations (5)–(7). The F-score ≥ 0.8 between two annotation results is considered to indicate acceptable consistency of the corpus.[29]

$$P=\frac{N(\mathrm{A1}_{\mathrm{result}}=\mathrm{A2}_{\mathrm{result}})}{N(\mathrm{A2}_{\mathrm{result}})} \qquad (5)$$

$$R=\frac{N(\mathrm{A1}_{\mathrm{result}}=\mathrm{A2}_{\mathrm{result}})}{N(\mathrm{A1}_{\mathrm{result}})} \qquad (6)$$

$$F = \frac{P \times R \times 2}{P + R} \tag{7}$$

where $A1_{result}$ is the annotation result of A1, $A2_{result}$ is the annotation result of A2, and N is the specific number of annotation result.

NER task consists of two subtasks, boundary detection and type identification.[18] Exact-match evaluation requires both the correct identification of the entity boundary and type.[30] In this study, the consistency of the entity was based on exact-match evaluation, which was referred to as entity-level IAA (2,2) or entity-level IAA (2,1). Moreover, the consistency of each Chinese character was evaluated using the BIOES annotation method (B-begin, I-inside, O-outside, E-end, S-single), called label-level IAA(2, 2), as shown in Appendix Table A.1. In particular, the consistency evaluation of a corpus is entity-based. The label-level IAA (2,2) only aims to obtain richer information to update the annotation principles and improve the entity-level IAA (2,2) score. Unless otherwise specified, both IAA (2,2) and IAA (2,1) refer to entity-based consistency evaluation.

The overall entity-level IAA (2,2) is provided in Table 1, while the entity-level IAA (2,2) of six types of entities is listed in Table 2. By introducing label-level IAA (2,2) information(shown in Figure 4), the entity-level IAA (2,2) reached 97.57% in fourth round. This shows that the introduction of label-level IAA (2,2) is effective in updating the annotation principles and enhancing entity-level IAA (2,2). The detailed discussions of label-level IAA (2,2) are presented in Appendix A.3. In addition, equations (5)–(7) were used to calculate IAA (2,1), where and $A2_{result}$ are the results of each repetition of [1×5 EMRs] for each annotator, and N is the total number of entities in the annotation result. The IAA (2,1) for A1 and A2 both exceeded 98% (shown in Figure 5). This shows that each annotator's labeling is stable and has a high standard.

Scale evaluation of corpus. In this study, the EMRs of 281 patients with CP were included. Their demographic information, including gender and age, is presented in Figure 6. As shown in the Figure 6(a), 64.06% were male and 35.94% were female. As shown in the Figure 6(b), the ages of patients were divided into five stages: (1) infancy, 0 to 1 years old, encompassing 0 to 12 months of age; (2) early childhood, 1 to 3 years old; (3) preschool, 3 to 6 years old; (4) school age, 6 to 12 years old; and (5) adolescence, 12 to 18 years old. It can be seen from the picture that children aged 1 to 3 years are the most common, followed by those aged 3 to 6 years.

Table 1. Entity-level inter-annotator agreement (IAA (2,2)) evaluation results ($A1_{result}$ is the annotation result of A1, $A2_{result}$ is the annotation result of A2, N is the specific number of annotation result; CA is the consistent entities between two annotators).

	IAA (2,2) (%)	$N(A1_{result})$	$N(A2_{result})$	N(CA)
Round 1	87.46	4277	4362	3778
Round 2	92.87	6569	4575	4246
Round 3	96.67	5075	5100	4918
Round 4	97.57	5124	5131	5003

In Table 3, six types of entities across five categories of EMRs were statistically analyzed after labeling 281×5 EMRs. The total number of entities in the corpus is 127,523, the number of unique entities is 24,424. Among the six types of entities, SYMPTOM accounts for the majority, while DISEASE and TREATMENT are relatively few in number. Among these five categories of EMRs, the *medical record characteristics* encompassed the greatest quantity of entities, whereas the *diagnostic basis* contained the least number of entities. Therefore, in the subsequent model-training phase, the five categories of EMRs from each patient were combined into training, validation, or test sets to prevent data imbalance. In Table 4, we also counted nested and long entities in the corpus, as identifying these two types of entities is particularly challenging with current NER technology. Nested entities comprise 24.99% of the corpus, while long entities—those exceeding ten Chinese characters—make up 5.98%. After entity de-duplication, the proportions change significantly: nested entities drop to 9.79%, and long entities increase to 19.25%. These differences highlight the impact of de-duplication on entity proportions.

Assessment results of NER model

To comprehensively evaluate the model's effectiveness, we compared the performance of our model with other state-of-the-art (SOTA) models including HMM,[31,32] CRF, BiLSTM, BiLSTM-CRF,[33,34] and BERT-BiLSTM-CRF[35–37] for both flat and nested entities in our corpus. The performances of flat entities are listed in the left side of Table 5, it is observed that, except for the lower F1-score of the HMM, the F1-score of the other models do not show significant differences. All of the models achieved excellent performance with F1-score exceeding 90%. The performances of nested entities are listed in the right side of Table 5. Our model identified nested entities significantly better than the other SOTA models, with a F1-score approximately 40% higher than the other models. This indicates that these advanced SOTA methods are suitable for the recognition of flat entities, but they are less effective in recognizing of nested entities in our corpus.

To further objectively assess the performance of our NER model from multiple dimensions, we performed an

Table 2. Entity-level inter-annotator agreement results for six types of entities.

	SYMPTOM (%)	BODY (%)	TREATMENT (%)	DISEASE (%)	CHECK (%)	ADJUNCT (%)	Average (%)
Round 1	95.00	80.79	89.11	69.84	78.76	74.66	87.46
Round 2	95.75	90.13	94.63	90.41	91.80	82.73	92.87
Round 3	98.93	96.02	95.32	92.47	94.49	92.02	96.67
Round 4	98.95	97.47	97.81	97.81	95.77	94.34	97.57

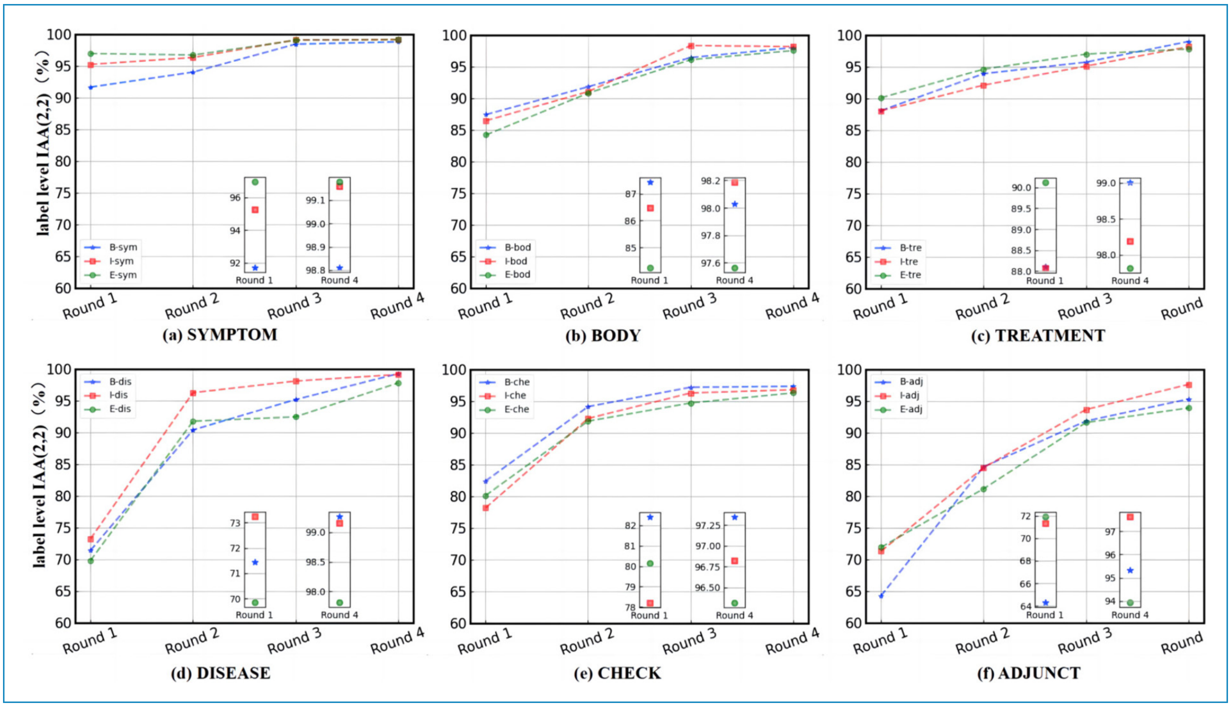


Figure 4. Label level inter-annotator agreement (IAA (2,2)) statistical analysis diagram of four rounds annotation of six types of entities. The label-level IAA (2,2) of all six types entities exhibited a rising trend, and the *B*-label, *I*-label, and *E*-label of each entity eventually tended to be high and consistent.

in-depth analysis of its performance across gender and age groups (Table 6), six types of entities (Figure 7) and five categories of EMRs (Figure 8). In Table 6, the performance of model across gender and age subgroups is consistent with its overall performance (Table 5), though the difference in performance between teenagers and non-teenagers is more pronounced than that between the gender subgroups. Our NER model generally performs well overall but has shown some shortcomings with certain types of entities or specific categories of EMRs. In Figure 7, both NER models perform poorly on the ADJUNCT type compared to other entity types. Additionally, both models exhibit relatively low *P*-score across all entity types. In Figure 8, both NER models perform the worst on *hospital course*. Similar to their performance on the six types of entities, both models exhibit relatively low *P*-value across all categories of EMRs.

Discussion

Entity corpus

We constructed a corpus of EMRs for children with CP under the guidance of clinical experts and combined it with clinical requirements under strict quality control. In this study, the IAA (2,2) between the two annotators was first tested across four rounds of labeling. Subsequently, the IAA (2,1) of each annotator was tested using test-retest reliability during the fourth round of formal labeling. And we introduce label-level IAA (2,2) to improve the

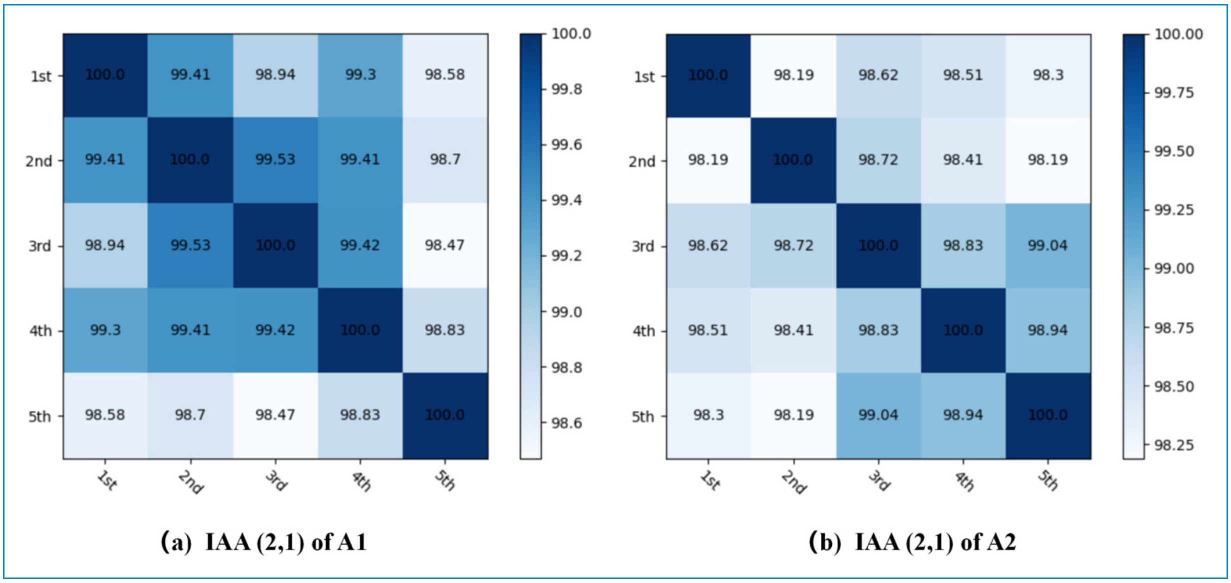


Figure 5. Intra-annotator agreement (IAA (2,1)) of entity-level. The horizontal and vertical axes of the confusion matrix depict the results of same [1 × 5 EMRs] labeled during the *n*-th repetition. Results of A1 are presented on the left and results of A2 are shown on the right.

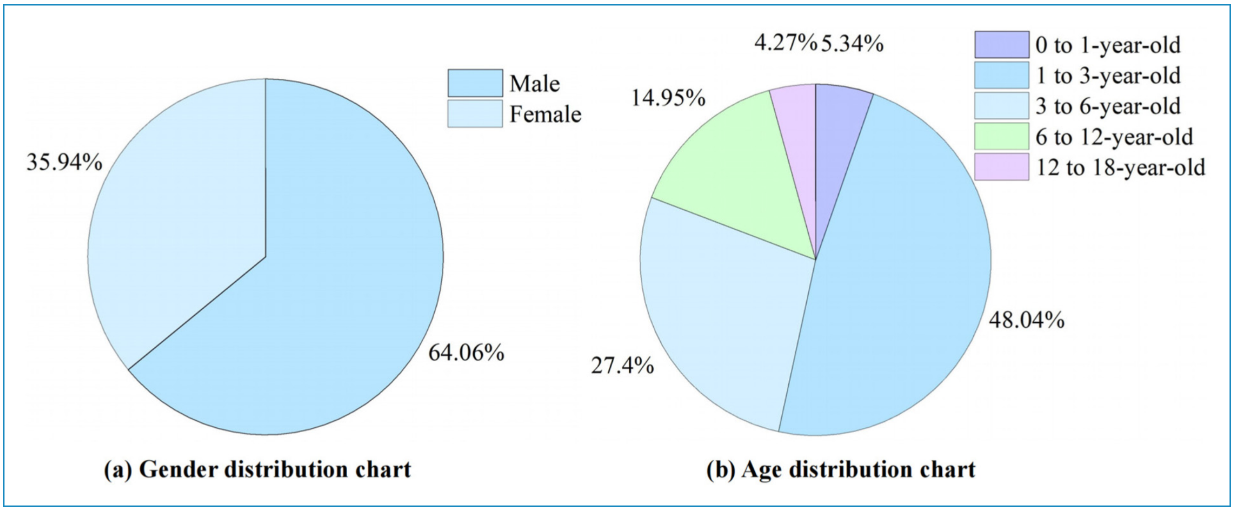


Figure 6. Demographic information of 281 cerebral palsy patients in this corpus.

annotation principle and enhance entity-level IAA (2,2). These approaches ensure more stringent quality control compared to other corpus builds that only calibrate entity-level IAA (2,2). The corpus demonstrated high agreement, as indicated by the IAA (2,2) and IAA (2,1), reflecting the confidence and reliability of the annotations. The corpus was sizable, comprising 1405 real EMRs of CP. The total number of annotated entities reached 127,523, with 24,424 unique entities, surpassing the size of current EMRs entity corpus.[8,10,11,20] Furthermore, during our analysis of patient demographics, we discovered that the documentation of Gross Motor Function Classification System (GMFCS) levels and CP classification was incomplete. This lack of thorough documentation could adversely affect future clinical studies. Therefore, we strongly recommend the adoption of standardized and structured documentation practices for medical procedures.

Performance analysis of NER model

In this study, we proposed the hypothesis that a NER model combining Chinese clinical BERT, the attention mechanism, and a GP network would achieve superior performance on CP EMRs, particularly in recognizing nested entities. In Table 5, our model, along with other SOTA models, achieves commendable performance on flat entity recognition tasks.

Table 3. Statistics on six types of entities across five categories of electronic medical records in the corpus.

	Medical history characteristics	Discharge summary	History of present illness	Diagnosis basis	Hospital course	Total
SYMPTOM	27,825(8975)	16,435(5027)	8219(3590)	3716(2077)	6437(3736)	62,632(15,808)
BODY	8005(874)	4674(421)	1550(278)	1153(327)	2040(660)	17,422(1396)
TREATMENT	1217(360)	400(42)	1213(347)	132(65)	3362(826)	6324(1176)
DISEASE	890(307)	53(27)	787(222)	1176(361)	182(100)	3088(643)
CHECK	10,731(832)	5297(238)	1433(212)	2278(329)	7095(933)	26,834(1584)
ADJUNCT	3890(1965)	1703(541)	2352(1153)	1157(361)	2121(1357)	11,223(4103)
Total	52,558(13,207)	28,562(6261)	15,554(5773)	9612(3515)	21,237(7574)	127,523(24,424)

The numbers in parentheses represent the unique entities after de-duplication.

Table 4. Statistics on nested and long entities in the corpus.

	Medical history characteristics	Discharge summary	History of present illness	Disgnosis basis	Hospital course	Total
Total	52,558(13,207)	28,562(6261)	15,554(5773)	9612(3515)	21,237(7574)	127,523(24,424)
Nested entities number	14,074(1425)	8400(701)	2358(492)	2394(611)	4639(922)	31,865(2390)
Nested entity ratio (%)	26.78(10.79)	29.41(11.20)	15.16(8.52)	24.91(17.38)	21.84(12.17)	24.99(9.79)
Long entities number	2447(1847)	839(629)	674(608)	643(456)	3027(2054)	7630(4701)
Long entity ratio (%)	4.66(13.99)	2.94(10.05)	4.33(10.53)	6.69(12.97)	14.25(27.12)	5.98(19.25)

A long entity is defined as one that exceeds ten Chinese characters in length. The numbers in parentheses represent either the unique entities after de-duplication or the proportions calculated from the unique entities.

However, when it comes to nested entity recognition, our model demonstrates a significant advantage over the others. The experimental results validate this hypothesis. This success is attributable to the strengths of each component within our model. Replacing BERT-base with bertcner in our model yields better results, demonstrating that Chinese clinical BERT is more effective for entity recognition in the Chinese clinical corpus. Pre-trained on a substantial amount of clinical data, Chinese clinical BERT provides a robust foundation for understanding the nuances of medical language and terminology. Our model (BERT-base), utilizing BERT-base as the embedding layer, demonstrates superior performance compared to the widely used BERT-BiLSTM-CRF (BERT-base) model. This improvement underscores the effectiveness of the multi-head attention mechanism and the GP network in recognizing nested entities. The attention mechanism captures rich contextual semantic information, while the two-dimensional span matrix of the GP network enables appropriate multi-label decoding of nested entities, further enhancing the overall performance. Specifically, for a sentence of length n, there can be up to $n(n+1)/2$ candidate entities if there is no restriction on entity length and entities are allowed to be nested within each other. Traditional one-dimensional models like HMM and CRF are limited in their capacity to decode only n labels (which corresponds to n labels, not n entities), take for example the flat entity sequence in Appendix Table A.1. Given that $n(n+1)/2 \gg n$, these models are

Table 5. Performance of our named entity recognition model and other SOTA models on flat entities and nested entities (P: precision; R: recall; F1: F1-score; our model: BERT + multi-head attention mechanism + span matrix of global pointer).

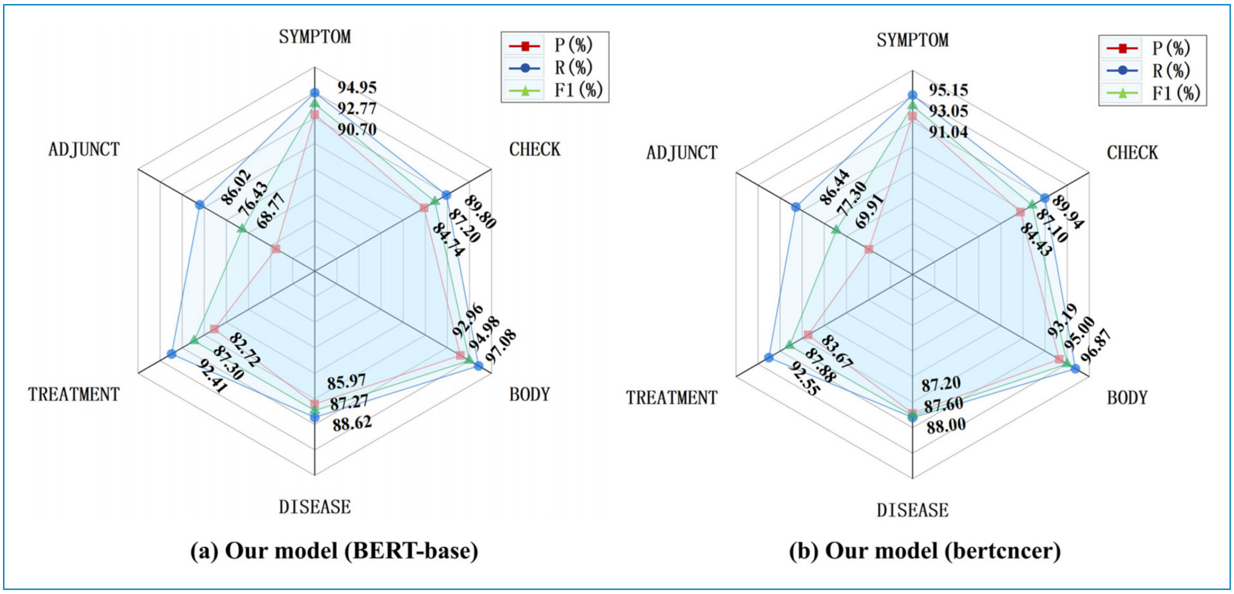
	Flat entities			Nested entities		
Model	*P* (%)	*R* (%)	F1 (%)	*P* (%)	*R* (%)	F1 (%)
HMM	79.91	83.74	81.56	39.60	62.42	45.91
CRF	96.05	94.93	95.46	48.12	51.60	51.60
BiLSTM	96.94	94.19	95.45	46.97	67.79	51.77
BiLSTM-CRF	96.17	95.56	95.85	48.54	70.08	52.96
BERT-BiLSTM-CRF (BERT-base)	95.11	95.71	95.41	44.33	21.86	29.28
Our model (BERT-base)	91.74	95.10	93.39	87.29	92.82	89.91
Our model (bertcner)	92.06	95.19	93.59	86.31	93.15	90.15

Table 6. Performance of the NER model across gender and age groups (P: precision; R: recall; F1: F1-score; our model: BERT + multi-head attention mechanism + span matrix of global pointer).

	Our model (BERT-base)			Our model (bertcner)		
	P (%)	*R* (%)	F1 (%)	*P* (%)	*R* (%)	F1 (%)
Male	86.33	92.56	89.33	86.59	92.46	89.43
Female	87.62	93.64	90.53	88.03	94.01	90.92
Teenager	83.60	91.71	87.47	85.01	91.57	88.17
Non-teenager	87.14	93.16	90.05	87.41	93.30	90.26

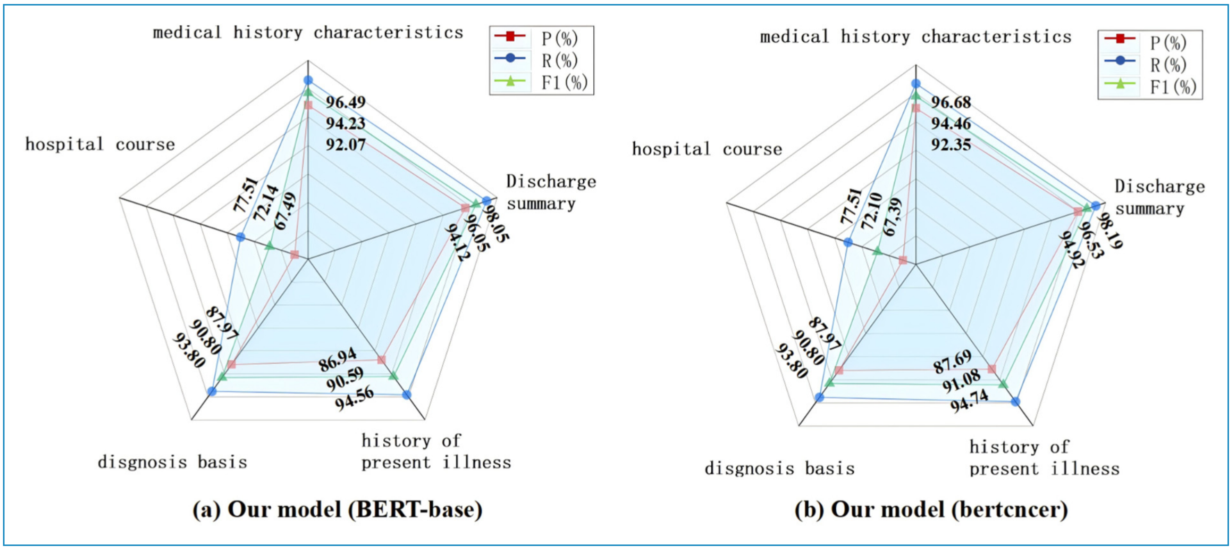
inherently inadequate for recognizing nested entities. In contrast, the GP constructs a $n \times n$ span matrix for each attention head (the *Tag Decoder* in Figure 3), with each head corresponding to a specific type of entity. This $n \times n$ configuration is sufficient to cover the $n(n+1)/2$ candidate entities, thereby enabling the effective decoding of nested entities. Consequently, the GP's architectural advantage lies in its ability to handle the complexity of nested entity structures, providing a robust solution where traditional models fall short.

The specific performance of our NER model is depicted in greater detail in Table 6, Figures 7 and 8. In Table 6, the performance of the model in terms of gender and age is basically consistent with the overall performance of the model (Table 5), demonstrating strong robustness across different demographic groups. In addition, there is a larger difference in performance between the age subgroups compared to the gender subgroups. This disparity is likely due to the relatively balanced gender distribution in the corpus (Figure 6), while the teenager group, representing only 4.27% of the total dataset, has less training data, leading to reduced performance in this subgroup. In Figure 7, both NER models perform poorly on the ADJUNCT type compared to other entity types. We speculate that there are two primary factors contributing to the lower F1-score observed in ADJUNCT. Firstly, the models, including BERT-base and bertcner, lack training data relevant to ADJUNCT entities, which diminishes their recognition capability for this type. Secondly, in Table 3, the total number of ADJUNCT is 11,223, and the number after de-duplication is 4103. The ratio before and after de-duplication is small compared to other types of entities, indicating that the same ADJUNCT occurs less frequently in the training data. In Figure 8, both NER models perform the worst on *hospital course*. As shown in Table 3, the number of de-duplication ADJUNCT in *hospital course* is 1,357, which is significantly higher than that of other categories of EMRs. In addition, in Table 4, the percentage of de-duplication long entities in *hospital course* is as high as 27.12%. We speculate that the large number of ADJUNCT and long entities in *hospital course* both contribute to the increased difficulty for the NER model. In both the six types of entities shown in Figure 7 and the five categories of EMRs in Figure 8, the *P*-score of model is relatively low. This suggests that the model likely recognizes many irrelevant entities. Upon conducting a comprehensive error analysis, the fact that the NER model incorrectly identifies redundant entities was confirmed, such as *one month of treatment in a local hospital* ADJUNCT entity is redundantly divided into nested entities



(a) Our model (BERT-base) **(b) Our model (bertcncer)**

Figure 7. Performance of our NER model on six types of entities (P: precision; R: recall; F1: F1-score).



(a) Our model (BERT-base) **(b) Our model (bertcncer)**

Figure 8. Performance of our NER model across five categories of EMRs (P: precision; R: recall; F1: F1-score).

such as *one month of treatment* and *a local hospital*. Improving the recognition accuracy of the ADJUNCT type and reducing the recognition of redundantly entities are the primary enhancement directions for future research.

Clinical and social potential

Children with CP and their families endure significant long-term rehabilitation challenges and economic pressures. The capacity for pediatric rehabilitation services in China remains inadequate, with a pronounced disparity in the distribution of resources. To address this dilemma, it is crucial to leverage extensive medical data for in-depth analysis to enhance rehabilitation medical services and improve the quality of life for these children. The EMRs of CP not only contain critical diagnostic and treatment information but also encapsulate the clinical reasoning of physicians, offering substantial knowledge reuse potential. Nevertheless, much of the information within EMRs exists as unstructured text, posing challenges for direct analysis and utilization. By extracting key information from EMRs, a knowledge base for motor function rehabilitation in CP can be developed. This knowledge base will encompass detailed data on medical histories, treatment plans, and rehabilitation outcomes, facilitating the systematic synthesis and dissemination of rehabilitation experiences. Integrated into a clinical decision support system, this knowledge base leverages massive case data to provide evidence-based recommendations, assisting clinicians in making more informed and precise decisions. For instance, the knowledge base enables physicians to efficiently identify similar cases and review their treatment plans, thereby crafting more personalized rehabilitation strategies for new patients and enhancing rehabilitation outcomes.

Conclusion

In this study, we successfully constructed a high-quality and comprehensive corpus of EMRs specifically for CP rehabilitation, meticulously annotated and rigorously evaluated for consistency. The incorporation of test-retest reliability in the annotation process underscores the confidence and reliability of the data. Building on this corpus, we developed a robust NER model using advanced deep-learning techniques, including Chinese Clinical BERT, multi-head attention, and the GP network. Our model excels in recognizing nested entities, enabling more nuanced and detailed analysis of CP patient data. This endeavor established a solid foundation of data and models for advanced mining and analysis of extensive CP EMRs. The results are expected to significantly advance knowledge representation and facilitate intelligent applications in related medical fields.

Limitations

We must acknowledge that our study has certain limitations. Future research could benefit from gathering a larger dataset of EMRs from multiple medical centers to further expand the corpus. Additionally, there is a need to focus on enhancing the performance of ADJUNCT type entities and *hospital course* category EMRs, which could improve overall model performance. Addressing these limitations will contribute to more robust and comprehensive findings in subsequent studies.

Code and data availability: The code and data used in this study are available upon request. To access the datasets, please contact the corresponding author at *gq_jin18@outlook.com*. We will provide the relevant materials to researchers for reproducibility and further analysis.

Contributorship: Meirong Xiao performed management activities to annotate, designing computer programs, mathematical formulations, and specifically writing the initial draft. Qiaofang Pang did testing of existing code components, and writing—reviewing and editing. Yean Zhu suggested ideas, specifically critical review. Shuai Lang did validation and supervision. Guoqiang Jin suggested ideas, specifically critical review, and did supervision.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Institutional Review Board of Children's Hospital of Chongqing Medical University (No. 2023242).

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the Jiangxi Province 03 Special and 5G Project (No. 20224ABC03A04), and in part by the Key Research and Development Program of Jiangxi Province (No. 20202BBE53021).

Guarantor: Yean Zhu.

ORCID iD: Yean Zhu https://orcid.org/0000-0003-1199-7099

References

1. McIntyre S, Goldsmith S, Webb A, et al. Global prevalence of cerebral palsy: a systematic analysis. *Dev Med Child Neurol* 2022; 64: 1494–1506.
2. Shengyi Y, Jiayue X, Jing G, et al. Increasing prevalence of cerebral palsy among children and adolescents in China 1988–2020: a systematic review and meta-analysis. *J Rehabil Med* 2021; 53: jrm00195.
3. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020; 396: 1204–1222.
4. Demont A, Gedda M, Lager C, et al. Evidence-based, implementable motor rehabilitation guidelines for individuals with cerebral palsy. *Neurology* 2022; 99: 283–297.
5. Ji B, Liu R, Li S, et al. A hybrid approach for named entity recognition in Chinese electronic medical record. *BMC Med Inform Decis Mak* 2019; 19: 149–158.
6. Lee L-H and Lu Y. Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition. *IEEE J Biomed Health Inform* 2021; 25: 2801–2810.
7. Song B, Li F, Liu Y, et al. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Brief Bioinform* 2021; 22: bbab282.
8. An Y, Xia X, Chen X, et al. Chinese Clinical named entity recognition via multi-head self-attention based BiLSTM-CRF. *Artif Intell Med* 2022; 127: 102282.
9. Bekteshi S, Monbaliu E, McIntyre S, et al. Towards functional improvement of motor disorders associated with cerebral palsy. *Lancet Neurol* 2023; 22: 229–243.
10. Gao Y, Gu L, Wang Y, et al. Constructing a Chinese electronic medical record corpus for named entity recognition on resident admit notes. *BMC Med Inform Decis Mak* 2019; 19: 67–78.

11. Fang A, Hu J, Zhao W, et al. Extracting clinical named entity for pituitary adenomas from Chinese electronic medical records. *BMC Med Inform Decis Mak* 2022; 22: 72.
12. O'Connor K, Sarker A, Perrone J, et al. Promoting reproducible research for characterizing nonmedical use of medications through data annotation: description of a Twitter corpus and guidelines. *J Med Internet Res* 2020; 22: e15861.
13. Leaman R and Lu Z. Taggerone: joint named entity recognition and normalization with semi-Markov models. *Bioinformatics* 2016; 32: 2839–2846.
14. Kaewphan S, Hakala K, Miekka N, et al. Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling. *Database* 2018; 2018: bay096.
15. Li P-H, Fu T-J and Ma W-Y. Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER. In: Proceedings of the AAAI conference on artificial intelligence, 2020, pp.8236–8244.
16. Huang Z, Xu W and Yu K. Bidirectional LSTM-CRF models for sequence tagging. Arxiv Preprint Arxiv:1508.01991, 2015.
17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30: 1706.03762.
18. Li J, Sun A, Han J, et al. A survey on deep learning for named entity recognition. *IEEE Trans Knowl Data Eng* 2020; 34: 50–70.
19. Devlin J, Chang M-W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NaacL-HLT, 2019, pp.2.
20. Li X, Zhang H and Zhou X-H. Chinese Clinical named entity recognition with variant neural structures based on BERT methods. *J Biomed Inform* 2020; 107: 103422.
21. Xie Q, Bishop JA, Tiwari P, et al. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowl Based Syst* 2022; 252: 109460.
22. Su J, Murtadha A, Pan S, et al. Global pointer: novel efficient span-based approach for named entity recognition. Arxiv Preprint Arxiv:2208.03054, 2022.
23. Uzuner Ö, South BR, Shen S, et al. 2010 I2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18: 552–556.
24. He B, Dong B, Guan Y, et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. *J Biomed Inform* 2017; 69: 203–217.
25. Bai W, Al-Karaghouli M, Stach J, et al. Test–retest reliability and consistency of HVPG and impact on trial design: a study in 289 patients from 20 randomized controlled trials. *Hepatology* 2021; 74: 3301–3315.
26. Su J, Lu Y, Pan S, et al. Roformer: enhanced transformer with rotary position embedding. Arxiv Preprint Arxiv:2104.09864, 2021.
27. Pan Y-C, Liu Y-Y and Lee L-S. Named entity recognition from spoken documents using global evidences and external knowledge sources with applications on mandarin Chinese. In: IEEE workshop on automatic speech recognition and understanding, 2005, 2005, pp.296–301: IEEE.
28. Su J, Zhu M, Murtadha A, et al. Zlpr: a novel loss for multi-label classification. Arxiv Preprint Arxiv:2208.02955, 2022.
29. Artstein R and Poesio M. Inter-coder agreement for computational linguistics. *Comput Linguist* 2008; 34: 555–596.
30. Peshterliev S, Dupuy C and Kiss I. Self-attention gazetteer embeddings for named-entity recognition. Arxiv Preprint Arxiv:2004.04060, 2020.
31. Li Y, Shetty P, Liu L, et al. BERTifying the hidden Markov model for multi-source weakly supervised named entity recognition. Arxiv Preprint Arxiv:2105.12848, 2021.
32. Li Y, Song L and Zhang C. Sparse conditional hidden Markov model for weakly supervised named entity recognition. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, 2022, pp.978–988.
33. Deng N, Fu H and Chen X. Named entity recognition of traditional Chinese medicine patents based on BiLSTM-CRF. *Wirel Commun Mob Comput* 2021; 2021: 1–12.
34. Cho H and Lee H. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinf* 2019; 20: 1–11.
35. Liu J, Gao L, Guo S, et al. A hybrid deep-learning approach for complex biochemical named entity recognition [Formula presented]. 2021.
36. Li W, Du Y, Li X, et al. UD_BBC: named entity recognition in social network combined BERT-BiLSTM-CRF with active learning. *Eng Appl Artif Intell* 2022; 116: 105460.
37. Zheng Y, Han Z, Cai Y, et al. An imConvNet-based deep learning model for Chinese medical named entity recognition. *BMC Med Inform Decis Mak* 2022; 22: 303.

Appendix A

A.1. Examples of BIOES labeling(B-begin, I-inside, O-outside, E-end, S-single)

Character sequence	表	现	为	足	跟	不	着	地	，	伴	双	下	肢	肌	张	力	增	高	。
	Performance for the heel cannot ground, lower limb muscle tension increased.																		
Entity tags sequence	O	O	O	B-sym	I-sym	I-sym	I-sym	E-sym	O	O	B-sym	I-sym	I-sym	I-sym	I-sym	I-sym	I-sym	I-sym	O
Nested annotation tag	O	O	O	B-bod	E-bod	O	O	O	O	O	B-bod	I-bod	E-bod	B-che	I-che	E-che	O	O	O

A.2. Annotation principles

Although we used the UMLS semantic types as a reference, the boundaries and types of entities showed slight variations. Six types—SYMPTOM, BODY, TREATMENT, DISEASE, CHECK, and ADJUNCT—were used in the annotation principles to classify the medical entities. The following examples consist of literal translations from Chinese to English.

SYMPTOM refers to the normal and abnormal states of the patient, including normal and abnormal findings obtained from examinations, as well as the symptoms or signs exhibited by the patient. Normal results were included as they hold significance in disease diagnosis. Which corresponds to a UMLS sign, symptom, test results, etc., and is marked as sym.

Example: 患儿存在[步态不稳]*sym*及[行走姿势异常]*sym*表现 (There were manifestations of [gait instability] *sym* and [abnormal walking posture]*sym*.)

BODY refers to the cells, tissues, organs, systems, and limbs of the human body, including human metabolic substances. Which corresponds to a UMLS body part, organ, and organ component and is marked as bod.

Example: [幕上脑室]*bod*积水，[中脑导水管]*bod*显示欠清，[后颅窝]*bod*囊肿 (hydrosis [supratentorial ventricle]*bod*, hypoclearness [midbrain aqueduct]*bod*, cyst [posterior fossa]*bod*.)

TREATMENT refers to therapeutic procedures, interventions, and medications administered to patients to address a disease or alleviate its symptoms. This corresponds to UMLS clinical drugs, antibiotics, therapeutics, preventive procedures and is marked as tre.

Example: 行[减重训练]*tre*，[沙盘游戏]*tre* (Do [weight loss training]*tre*, [sandplay]*tre*.)

DISEASE refers to the cause of a patient's unhealthy state or a diagnosis made by a physician, including a diseases, syndromes, poisoning, injuries, organ damage, or cell damage. This corresponds to UMLS diseases, syndromes, injuries, and poisoning and is marked as dis.

Example: 出生时存在[缺血缺氧性脑病]*dis*基础疾病 (There was a underlying disease at birth [hypoxic ischemic encephalopathy]*dis*.)

CHECK refers to the process of conducting examinations on a patient to identify or confirm a disease or symptom, as well as gather more information about the disease or symptoms. This corresponds to a UMLS laboratory procedure, laboratory test, and diagnostic procedure and is labeled as che.

Example: [头颅核磁共振成像]*che*未见明显异常 (There was no significant abnormality on [head Magnetic Resonance Imaging]*che*.)

ADJUNCT includes seven modifiers of diseases or symptoms (absent, family, present, conditional, possible, hypothetical, and occasional), three modifiers of treatments (history, absent, and present), and all other entities describing modifying functions, which are simply labeled as ADJUNCT classes and marked as adj.

Example: 右手可伸手抓物，但[欠灵活]*adj* (The right hand can reach out to grasp, but [lacks dexterity]*adj*.)

General principles The general boundaries and types of medical entities are described above. The following principles should also be followed in the actual labeling.

1. Conjunctions and punctuation should be minimized within the entity as much as possible.

   Example: [血常规]*che*及[大小便常规]*che*未见明显异常 (There were no obvious abnormalities in [blood routine]*che* and [urine routine]*che*.) (True)

   [血常规及大小便常规]*che*未见明显异常 (There was no obvious abnormality in [blood routine and urine routine]*che*.) (False)
2. Abbreviations, Chinese acronyms, or common names within the entity should be appropriately marked or indicated.

   Example:入院查体：[T]*che* 36.5°C，[P]*che* 32次/分，[R]*che* 103次/分 (Admission for physical examination: [T]*che* 36.5°C, [P]*che* 32times/min, [R]*che* 103 times/min.)
3. We only allowed the nested annotation of SYMPTOM entities. That is, if BODY, TREATMENT, DISEASE, or CHECK exist inside the SYMPTOM entity, the annotator should also nest-annotate them(Example 1). If the ADJUNCT in the SYMPTOM only modifies this SYMPTOM, it will not be nested(Example 2), unless the ADJUNCT also modifies other SYMPTOMs(Example3).

   Example1:[[双下肢屈肌]*bod*[肌张力]*che*增高]*sym* ([increased [muscle tone]*che* in [flexor muscles of both lower limbs]*bod*]*sym*)

   Example2:整个病程中[无发育倒退]*sym*，[无抽搐发作]*sym* ([no developmental regression]*sym*, [no convulsive seizures]*sym* throughout the course of the disease.)

   Example3:[[无]*adj*特殊气味]*sym*及[毛发变浅]*sym* ([[no]*adj* special odor]*sym* or [lightened hair]*sym*.)
4. All entities should comply with the maximum labeling principle. That is, if an entity contains other smaller entities, the largest entity should be labeled.

   Example:[右下肢屈肌]*bod* ([right lower limb flexors]*bod*) (True)

   [右下肢]*bod*[屈肌]*bod*([right lower limb]*bod*[¯exors]*bod*) (False)

A.3. Analysis of label-level IAA(2,2)

Table 2 shows that, after the first round of pre-annotation, CHECK, DISEASE, and ADJUNCT had an IAA(2,2) below 80%, even though the average IAA(2,2) of the six types of entities was greater than 80%. We employed the label-level IAA(2,2) to obtain more detailed information and enhance the entity-level IAA(2,2), as shown in Figure 3. We expect to maintain B-label, I-label, and E-label at high and consistent ideal levels for each entity, respectively. In the first round shown in Figure 3(a), the label-level IAA(2,2) of B-sym was lower than that of E-sym. The characters at the beginning of SYMPTOM, such as “[有反复痉挛样发作]*sym*” ([have repeated spasmoid attacks]*sym*) (The underline is the difference between A1 and A2, the square brackets in this paragraph indicates the correct boundary of the entity.), “[日间无发热]*sym*” ([during the day no fever]*sym*), and “[兴奋时有[左上肢]*bod*强直]*sym*” ([[left upper limb]*bod* stiffness when excited]*sym*) were found to be inconsistent after comparison, which requires further clarification and refinement of the upper boundary of SYMPTOM in the annotation principles. In the first round shown in Figure 3(b), B-bod, I-bod, and E-bod exhibited declining trends. We can assume that the two annotators have an inconsistent understanding of the lower boundary of BODY. After comparison, the annotator occasionally mislabeled the entities that were similar to “[双下肢屈肌]*bod*” ([both lower - limb flexor muscles]*bod*) as “[双下肢]*bod*"([both lower-limb] *bod*) and “[屈肌]*bod*” ([flexor muscles]*bod*), which is not in accordance with the maximum labeling principle (the fourth rule in the General principles in Appendix A.2). Therefore, we should label it as the “[双下肢屈肌]*bod*” ([both lower limb flexor muscles]*bod*), as well as the other entity types. In the first round shown in Figure 3(c), there is an inconsistent labeling, such as “行[肉毒素注射]*tre*” (do [botox injection]*tre*), and the B-tre is somewhat lower than the E-tre. Therefore, a further refinement of the upper boundary of TREATMENT is required. In the first round presented in Figure 3(d), the label-level IAA(2,2) of all DISEASE labels was considerably low. After comparison, we found that the classification between some DISEASE and SYMPTOM are not clearly defined, such as “[脑血管发育畸形]*dis*” ([cerebrovascular development malformation]*dis*), which is mislabeled as SYMPTOM. In the first round of Figure 3(e), I-che exhibits the lowest label-level IAA(2,2), which can be caused by the high percentage of CHECK with only two Chinese characters in this round, such as “[体温]*che*” [body temperature]*che* and “[心率]*che*” [heart rate] *che* after the conversion of the BIOES annotation method with labels only B-che and E-che. Therefore, we focused on analyzing B-che and E-che. E-che is relatively lower, such as “[头颅MRI检查]*che*” ([head MRI assessment]*che*) and “[粗大运动功能评估]*che*” ([gross motor function evaluation]*che*). This means that the lower boundary of CHECK requires further refinement. In the first round shown in Figure 3(f), the label-level IAA(2,2) for all ADJUNCT labels was low. We found that ADJUNCT is often labeled as non-entity in various situations, such as “[范

围较前扩大]*adj*” ([range expanded]*adj*), which is mislabeled as O. Moreover, B-adj is lower than E-adj, indicating that there are more upper boundary inconsistencies in ADJUNCT, such as “于[当地医院] *adj*” (at [local hospital]*adj*) and “[约入院前1+年]*adj* “” ([approximately 1 + years before admission]*adj*).

Based on the above analysis, we updated the annotation principles using medical specialists. The second round annotation results showed that annotation consistency improved significantly. However, we found that the label-level IAA(2,2) of B-sym, E-che, and E-adj was relatively low. Therefore, we used the same analysis and improvement methods as in the first round to make improvements and obtain better results in the third round compared to the second round. After several iterations of improvement, the results of the fourth round showed that B-label, I-label, and E-label of each type of entity were already at a high and consistent level. Eventually, the entity-level IAA(2,2) went from 87.46% in the first round to 97.57% in the fourth round, an improvement of approximately 10%. This shows that adding label-level IAA(2,2) information is convenient and effective for improving the annotation principles and enhancing entity-level IAA(2,2).