Original Paper

Symptom Terminology Normalization in Traditional Chinese Medicine: Development and Evaluation of a 2-Stage Deep Learning Framework Based on Fine-Grained Semantic Classification

Junyu Yao[1], MCM; Xingyue Gou[1], MMed; Wei Lai[1], PhD; Yuzhu Gao[1], MCM; Siqi Wang[2], MMed; Chuangan Zhou[1], BMed; Hui Ye[1], MEng; Jing Tian[3], MA; Jun Yi[4], MEng; Dong Cao[1], PhD

[1]School of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong, China
[2]Yunkang School of Medicine and Health, Guangzhou Nanfang College, Guangzhou, Guangdong, China
[3]School of Foreign Studies, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong, China
[4]School of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou, Guangdong, China

Corresponding Author:

Dong Cao, PhD
School of Medical Information Engineering
Guangzhou University of Chinese Medicine
232 Outer Ring East Road, Guangzhou University City, Panyu District
Guangzhou, Guangdong 510006
China
Phone: 86 13246852860
Email: caodong9@163.com

Abstract

Background: Due to the heterogeneity of symptom terminology and the lack of industry standards, the same symptom is often described using multiple expressions. Current normalization approaches struggle to comprehensively retrieve standard terms when a raw term maps to multiple symptoms.

Objective: This study aimed to address the lack of industry standards for traditional Chinese medicine (TCM) symptom terminology. This study proposed the split-then-concatenate normalization framework (STC-NF), a novel approach based on fine-grained semantic classification and a 2-stage deep learning architecture that uses electronic medical records (EMRs) as the data source.

Methods: This study proposed a 2-stage deep learning framework, "split-then-concatenate." In the splitting stage, TCM symptom entities were categorized into 12 fine-grained semantic labels, and 3 named entity recognition (NER) models were trained to extract TCM symptom terminology from EMRs. In the concatenation stage, standard terms with the same concept as raw terms were identified using a Bidirectional Encoder Representations from Transformers (BERT)–based binary classification model. The standard terms with specific semantic labels were concatenated and reordered according to predefined rules to output structured text, thereby normalizing TCM symptom terminology.

Results: The proposed STC-NF model achieved an accuracy of 91.4% (180/197) and an F_1-score of 360 out of 389 (92.5%) on the single-implication test set. For multi-implication terms, STC-NF achieved an accuracy of 84.3% (311/369) and an F_1-score of 1958 out of 2316 (84.5%), outperforming sequence generation in accuracy by 33.1 percentage points. On the mixed test set containing both single- and multi-implication terms, STC-NF achieved an accuracy of 88.1% (990/1124) and an F_1-score of 3862 out of 4385 (88.1%), exceeding the best-performing baseline model, multi-task candidate generator (MTCG), by 16.7 and 22.2 percentage points in accuracy and F_1-score, respectively.

Conclusions: In this study, we verified that the fine-grained semantic classification and the 2-stage "split-then-concatenate" framework effectively improved performance of named entity recognition and entity alignment, providing an improved approach to normalizing TCM symptom terminology.

JMIR Med Inform 2026;14:e85825; doi: 10.2196/85825

Keywords: symptom terminology; traditional Chinese medicine; natural language processing; named entity recognition; named entity normalization; pretrained language model

Introduction

Background and Rationale

Symptoms are the core basis of traditional Chinese medicine (TCM) identification and treatment. They are also one of the main input features for AI to realize assisted diagnosis and treatment [1], which are of great value to medical research [2]. Due to the heterogeneity of symptom terminology, the same symptom is often described in multiple expressions. For example, "失眠 (insomnia)," "不寐 (sleeplessness)," and "入睡困难 (difficulty initiating sleep)" are all described as the same clinical phenomenon in TCM electronic medical records (EMRs). In recent years, it has been widely acknowledged that normalizing symptom terminology is crucial for improving the accuracy of TCM syndrome differentiation and promoting the modernization of TCM [3]. However, neither industry consensus nor a national standard for TCM symptom terminology currently exists [4]. This gap has significantly hindered the sharing of medical data, knowledge mining, and the development of intelligent applications [5]. Efficiently and accurately aligning nonnormalized symptom raw terms with standard terms in the knowledge base has become a key issue to address.

Related Work

A number of recent studies have specifically targeted the normalization of TCM symptom terminology. Jia et al [6] proposed a symptom terminology normalization approach with hierarchical semantics (symNormHS) to address issues related to synonymous expressions and homographs through a 2-step process. The approach first extracted hierarchical semantic information from symptom terms using a multilabel text classifier to filter the candidate set. Then, it used a hybrid multigranularity text matching model, combined with an attention mechanism, to compute the similarity between raw terms and standard terms. Zhan et al [7] used an autostacker classifier model based on the automated machine learning (AutoML) framework to predict categories of TCM symptom terms, and computed semantic similarity between raw terms and standard terms using a directional skip-gram (DSG) model. Tang et al [8] retrieved candidate standard terms from Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) using a hybrid retrieval strategy that combined dice coefficients and term frequency-inverse document frequency (TF-IDF). The retrieved candidates were then reranked using enhanced representation through knowledge integration (ERNIE), specifically the ERNIE-Health model (Baidu Inc), which served as the semantic similarity scoring module for selecting the most semantically compatible standard terms. Zhou et al [9] unified the synonymous raw terms into standard terms by constructing a classification model for synonymous term conversion with the thinking of candidate terms (STC-TC), and misclassification detection was performed using the output comparison of multiple heterogeneous models (OCMH). Hu et al [10] proposed an approach to normalize TCM symptom terminology using Bidirectional Encoder Representations from Transformers (BERT)–bidirectional long short-term memory (BiLSTM)–conditional random field (CRF) and a multilabel classification strategy. By constructing an ontology framework for TCM symptoms and integrating knowledge graphs, this approach combined the correlated feature fusion module (CFFM) and the hierarchical labeling tree (HLT) to optimize the recognition and normalization capability of entities. The normalization approaches in [6-10] are all based on deep learning. Their innovations and limitations are summarized in Table 1.

Table 1. Innovations and limitations of different approaches.

Author	Year	Approach	Innovations	Limitations
Jia et al [6]	2021	symNormHS[a]	Enhanced semantic matching by fusing unigram- and bigram-level local features through the attention mechanism.	Low recall of multi-implication terms[b] (eg, "面红目赤 [flushed face with bloodshot eyes]" only matched "目赤 [bloodshot eyes]"), which required improving the normalization logic by formulating additional heuristic rules.
Zhan et al [7]	2021	DSG[c]	Constructed a classification model to handle symptom terms using the AutoML[d] framework and used the DSG model to capture the semantic relationships between raw terms and standard terms.	The DSG model had a limited ability to recognize multi-implication terms (eg, "头胀痛伴耳鸣如蝉 [distending headache with tinnitus like cicada chirping]"), which might lead to normalization failure due to excessively long text and semantic nesting.
Tang et al [8]	2022	ERNIE[e]-Health (Baidu)	Proposed a hybrid recall strategy (dice coefficient+TF-IDF[f]) to improve the recall of standard terms.	Unable to accurately deal with multi-implication terms (original description: 2 different concepts in one word), it was suggested to introduce external information, such as "context and symptomatic site," to address this issue.
Zhou et al [9]	2023	STC-TC[g]	Used the OCMH[h] to identify normalization errors.	Multi-implication terms (eg, "痰量多色白 [increased white-colored sputum]") had lower normalization accuracy due to their extremely low frequency of occurrence and the lack of sufficient training samples for the model to learn entity recognition laws.

Author	Year	Approach	Innovations	Limitations
Hu et al [10]	2024	Ontology-enhanced multilabel model	Introduced the CFFM[i] module to fuse entity features and the hierarchical labeling tree to solve the label imbalance problem, thereby enhancing normalization capability for rare symptoms.	The multilabel classification approach was unable to handle multi-implication terms composed of multiple symptom entities (eg, "腰膝酸软 [soreness and tenderness in the lower back and knees]"), as it only matched "腰酸软 (lower back soreness and tenderness)" and could not retrieve "膝酸软 (knee soreness and tenderness)."

[a]symNormHS: symptom terminology normalization approach with hierarchical semantics.
[b]In this study, we define a raw term that maps to exactly one standard term as a "single-implication term," and one that maps to multiple standard terms as a "multi-implication term."
[c]DSG: directional skip-gram.
[d]AutoML: automated machine learning.
[e]ERNIE: enhanced representation through knowledge integration.
[f]TF-IDF: term frequency-inverse document frequency.
[g]STC-TC: synonymous term conversion with thinking of candidate terms.
[h]OCMH: output comparison of multiple heterogeneous models.
[i]CFFM: correlated feature fusion module.

Beyond these TCM-specific systems, research on medical terminology normalization has more broadly progressed through 3 methodological stages, including rule-based and dictionary-matching approaches, machine learning approaches, and deep learning approaches. Rule-based and dictionary-matching approaches [11-16] require experts to manually construct specific regulations based on data features and then use specialized domain dictionaries to complete tasks such as named entity recognition (NER) and named entity normalization (NEN). Although such approaches achieve high accuracy when the lexicon is comprehensive, their performance is contingent on labor-intensive, manually crafted rules. Furthermore, reusing identical rule sets in new domains markedly reduces effectiveness, rendering the approach costly, laborious, and poorly portable. Machine learning approaches [17-22] can automatically extract features from medical corpora and predict entity classification labels through statistical computation, thereby eliminating the need to construct rules and effectively reducing manual effort. However, this approach computes the similarity of concepts by comparing the literal differences between 2 entities, which does not fully use the semantic information in the context, resulting in obvious limitations in its generalizability. Deep learning approaches [6-10,23-32] now dominate medical terminology normalization because they can embed phrases in dense vector spaces and capture latent semantic relations beyond surface forms. Recent studies have further shown that medical NER and normalization are affected by limited annotated data, entity-boundary ambiguity, heterogeneous preprocessing procedures, and insufficient use of contextual information. For Chinese clinical NER, Tang et al [33] proposed a segmentation synonym sentence synthesis mechanism to expand training data and improve model generalization. Li et al [34] explored a few-shot medical NER framework that combines retrieval-based example selection with a structured reasoning process, highlighting the importance of informative examples and task-specific guidance under limited annotation conditions. From the perspective of biomedical entity linking, Garda et al [35] emphasized that inconsistent preprocessing, knowledge-base selection, and evaluation protocols can limit reproducibility and comparability across normalization systems. Luo et al [36] further demonstrated that contextual information around entity mentions is important for biomedical named entity normalization, especially when mentions are semantically ambiguous. These studies collectively indicate that accurate clinical text standardization requires not only semantic matching but also reliable entity-boundary recognition, task-specific preprocessing, and context-aware alignment. To address these requirements from different methodological perspectives, existing deep learning approaches can be broadly categorized into 3 types based on their entity alignment mechanisms, including multiclassification, sequence generation, and binary classification.

The multiclassification paradigm casts normalization as a sentence-level classification task. Representatively, Huang et al [30] proposed a pretrained language model framework for *International Classification of Diseases* coding (PLM-*ICD*), which integrates domain-specific pretrained language models with the attention mechanism and thereby exploits contextual semantics to align uncommon raw terms (eg, "耳朵有蝉叫声 [ears buzzing]") with standard terms of low lexical overlap (eg, "耳鸣 [tinnitus]"). Compared with statistical models such as TF-IDF, Med, and BM25, which rely on character-level similarity, this paradigm attains markedly higher accuracy on single-implication terms; however, because its classification head emits only a single prediction for each input sentence, it cannot recover the multiple standard terms that a multi-implication expression requires. To overcome this single-output limitation, the sequence generation paradigm reformulates normalization as a token-by-token generation process. Along this line, Yan et al [31] generated standard terms sequentially and were thus able to align simply structured multi-implication terms that contain no overlapping, discontinuous, or nested entities (eg, "头昏嘴斜 [dizziness and mouth skewed]") with several standard terms (eg, "头晕 [dizziness]" and "口歪 [mouth skewed]"), thereby partially improving the accuracy of multi-implication normalization. Decomposing a raw term character by character, however, can fracture the structural integrity of an entity and degrade performance on single-implication terms. The binary classification paradigm instead preserves each candidate term intact and judges semantic equivalence in a pairwise manner. For

example, Liang et al [32] proposed the multi-task candidate generator (MTCG) model, which first retrieves the standard terms semantically closest to a raw term by computing the Euclidean distance between their vector representations, and then formats each raw term together with its retrieved candidates as sentence-pair inputs to a pretrained BERT model for binary classification; an output of 1 indicated semantic equivalence between the raw term and a candidate standard term, and the normalization of a multi-implication term is completed by aggregating all candidates judged to be equivalent.

Challenges in TCM Symptom Terminology Normalization

While the 3 paradigms reviewed above each advance entity alignment, none was designed to fully decompose the structurally complex symptom expressions that are common in TCM EMRs. As shown in Table 1, deep learning approaches can retrieve standard terms with low character-level similarity to raw terms—for instance, mapping "脑袋疼 (nǎo dai téng)" to "头痛 (tóu tòng)," both meaning "headache"—by capturing linguistic features and modeling contextual dependencies. However, for multi-implication terms composed of multiple symptom entities, current normalization approaches still perform with low accuracy. To better understand this limitation, after manually annotating 500 TCM EMRs, we found that multi-implication terms related to the TCM symptom terminology could be categorized into the following three types.

1. Overlapping entities: two symptom terms share the same suffix. For instance, the phrase "饮食睡眠尚可 (eating and sleeping okay)" is a combination of the symptom terms "饮食尚可 (eating okay)" and "睡眠尚可 (sleeping okay)."
2. Discontinuous entities: common TCM symptom terminology, such as "头痛 (headache)," "手麻 (hand numbness)," and "耳鸣 (tinnitus)," typically consists of an anatomy entity and an abnormality entity. However, sometimes the abnormality entity does not follow the anatomy entity. For instance, in the phrase "头有点痛 (a mild headache)," the adjective "有点 (mild)" (indicating symptom severity) separates the abnormality entity "痛 (pain)" from the anatomy entity "头 (head)."
3. Nested entities: two symptom terms share the same prefix. For instance, the phrase "尿频量多 (polyuria with increased urinary frequency)" results from combining symptom terms "尿频 (frequent urination)" and "尿量多 (copious urination)."

These 3 complex structural types are the main factors responsible for the low accuracy of current normalization approaches in handling multi-implication terms. To improve normalization performance for such terms, this study proposes the split-then-concatenate normalization framework (STC-NF), an approach based on fine-grained semantic classification that uses TCM EMRs as the data source.

Study Objective and Contributions

Guided by the objective of accurately normalizing such multi-implication terms, this study makes the following two principal contributions:

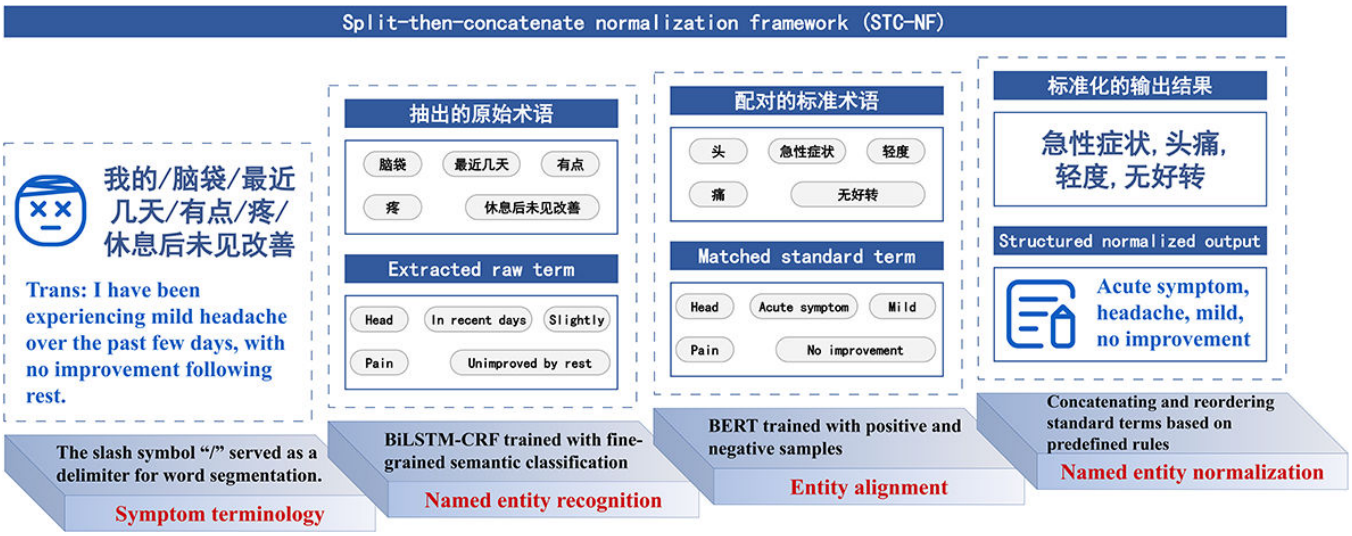
1. Fine-grained semantic classification rules are proposed to categorize entities related to TCM symptoms into 12 different types. Validation on 500 TCM EMRs demonstrates that labeling data with fine-grained semantic classification enables the NER model to achieve excellent performance. This approach could effectively address the difficulty in segmenting TCM symptom terminology caused by fuzzy entity boundaries.
2. Concatenation and ordering rules for standard terms are constructed to retrieve standard terms that semantically match the input raw term from the knowledge base more comprehensively when the latter maps to multiple standard terms. This approach shows excellent performance in handling multi-implication terms and could be applied to other medically relevant natural language processing (NLP) tasks, such as normalizing diagnoses or surgical procedures.

Methods

Overview of the Study

To accurately extract all symptom terms from EMRs and convert them into structured text, we proposed the STC-NF. This 2-stage process was designed after reviewing 3 normalization approaches, including multiclassification, sequence generation, and binary classification. In the first stage, we segmented sentences and extracted all symptom-related raw terms using fine-grained entity classification rules and a BiLSTM-CRF model. In the second stage, our study used a binary judgment model based on BERT to identify the conceptually equivalent standard term from a set of standard terms that shared the same semantic classification label as the raw term. Finally, we concatenated and reordered all standard terms according to predefined rules to generate the structured text, thereby completing the normalization of the entire sentence. The technical workflow of this study is shown in Figure 1.

Figure 1. Technical roadmap for normalization of traditional Chinese medicine (TCM) symptom terminology. BERT: Bidirectional Encoder Representations from Transformers; BiLSTM: bidirectional long short-term memory; CRF: conditional random field.



Introduction to Dataset Sources and Fine-Grained Semantic Classification

The primary source of the dataset in this study was the collection of 500 EMRs from the endocrine department of a hospital in Guangzhou, China. Before annotation, personally identifiable information was removed, and the free-text symptom descriptions were reviewed at the sentence level. Because this study used manually annotated free-text symptom mentions rather than structured laboratory or demographic variables as model inputs, conventional missing-value imputation was not performed. Sentences within the EMRs that lacked analyzable symptom descriptions were excluded from training instance construction; this sentence-level filtering did not affect the EMR-level dataset partitioning described below. The main features used in the NER stage were entity boundaries and fine-grained semantic labels, while the features used in the NEN stage were raw-term and candidate-standard-term pairs constrained by semantic label type. This design is consistent with recent clinical NER and biomedical entity-linking studies emphasizing task-specific annotation rules, limited-resource data construction, and standardized preprocessing for reproducible normalization experiments [33-35]. We categorized the symptom-related entities into 12 different types, and the specific classification rules and entity examples were as follows:

1. Independent symptom (label: Symp_S): a highly condensed name for a disease or an expression describing a preference, usually consisting of 2 Chinese characters. Examples: "咳嗽 (cough)," "呕吐 (vomiting)," "发热 (fever)," and "喜热饮 (preference for hot drinks)."
2. Negative occurrence (label: Neg): a character or word that describes a negative meaning. Examples: "无 (none)" and "未见 (not seen)."
3. Scope space (label: Pos_SCP): describes the scope in which the state occurs. Examples: "左 (left)," "右 (right)," "双侧 (bilateral)," and "小 (the character '小 [xiǎo]' in the TCM symptom terminology '小便 [xiǎo biàn]' implies 'a relation to urination or the urinary system')."
4. Primary space (label: Pos_Pri): describes the primary space in which the state occurs. Examples: "头 (head)," "神志 (mental)," "睡眠 (sleep)," and "便 (poop)."
5. Secondary space (label: Pos_Sub): after splitting the compound entity, describes a more specific level of the state. Examples: "量 (volume)" and "次数 (frequency)" in "小便量次数增多 (increased urinary frequency with elevated voided volume)."
6. Time of occurrence (label: Time): describes the time of occurrence of all abnormal symptoms. Examples: "三个月前 (3 months ago)" and "昨日 (yesterday)."
7. Occurrence condition (label: Cond): describes the specific time or condition under which an independent symptom occurs. Examples: "行走时 (when walking)," "平躺时 (when lying down)," "晨起时 (when waking up in the morning)," and "上楼后 (after going upstairs)."
8. Frequency of occurrence (label: Freq): describes how frequently the symptom occurs. Examples: "一日三次 (3 times a day)," "反复 (recurrent)," and "偶尔 (occasional)."
9. Qualitative severity (label: Sev_Qual): describes the qualitative seriousness of the symptom. Examples: "轻度 (mild)," "中度 (moderate)," and "重度 (severe)."
10. Quantitative severity (label: Sev_Quant): describes the quantitative severity of the symptom quantitatively. Examples: "约 50 斤 (about 25 kg)" in "体重下降约 50 斤 (significant body weight reduction of approximately 25 kg)."
11. Occurrence state (label: State): describes the abnormal manifestation of the symptom. Examples: "痛 (pain)," "刺痛 (stabbing pain)," "闷 (stuffiness)," "胀 (bloating)."
12. Occurrence trend (label: Trend): describes the trend in symptom change. Examples: "加重 (aggravation)," "减轻 (remission)," "缓解 (alleviation)."

All 500 EMRs were annotated according to these 12 fine-grained categories before model training and evaluation. The dataset was divided at a ratio of 8:2. A total of 400 EMRs, accounting for 30,490 out of 38,486 entities (79.2%), were used as the training set, and 100 EMRs, accounting for 7996 out of 38,486 entities (20.8%), were used as the test set.

Overview of the Entity Alignment Task

Entity alignment for symptom terminology aims to map each fine-grained raw symptom term extracted from medical texts to a semantically equivalent standard term in a standard knowledge base. Given a set containing n raw terms:

$$M = \{m_1, m_2, \cdots, m_n\} \quad (1)$$

and a standard knowledge base (containing *g* standard terms):

$$E = \{e_1, e_2, \cdots, e_g\} \quad (2)$$

The task is to map each fine-grained raw term m_i to a semantically equivalent standard term e_j in the standard knowledge base, with the candidate standard terms restricted to those sharing the same fine-grained semantic label as the raw term.

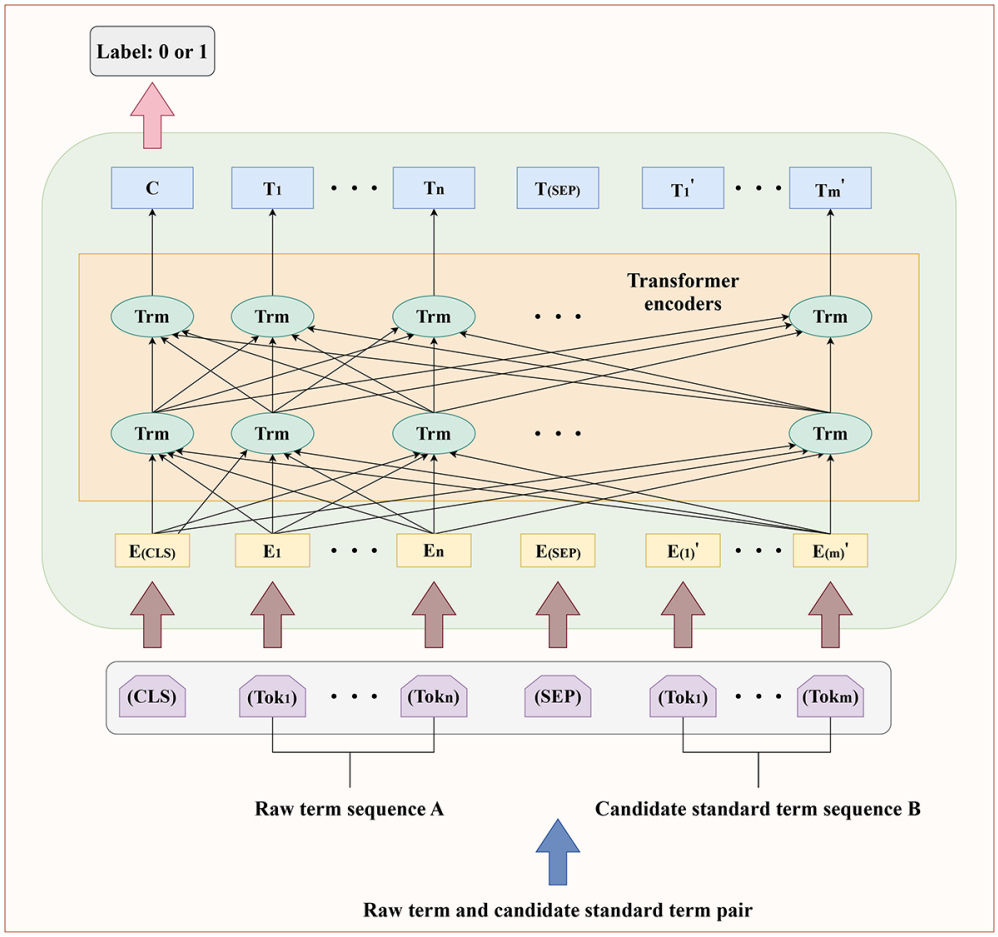
The core methodology involves training a deep semantic matching model, leveraging BERT's binary classification capability, to perform the normalization shown in Equation 3:

$$\text{norm}(m_i) = e_j \quad (3)$$

For a multi-implication symptom term, this mapping is applied to each of its extracted fine-grained raw terms, and the matched standard terms are subsequently concatenated and reordered according to predefined rules.

As illustrated in Figure 2, our proposed model processes a pair consisting of a raw term and a candidate standard term. The (CLS) token is placed at the beginning of the input sequence for the final text classification task, while the (SEP) token is used to separate the 2 terms. The model outputs a binary classification result: a label of "1" indicates a conceptual match, signifying that the candidate standard term is the correct alignment for the given raw term. Conversely, a label of "0" denotes a mismatch, meaning the candidate term is not the correct normalization. In such cases, the model iteratively evaluates the raw term against other candidate standard terms until a positive match (a result of "1") is identified.

Figure 2. Architecture of a Bidirectional Encoder Representations from Transformers (BERT)–based model for entity alignment. CLS: classification token; SEP: separator token; Tok: token; Trm: transformer module.



Introduction to the Methodology for Constructing the Standard Knowledge Base

Most of the standard terms we collected originate from the most widely used medical classifications and databases worldwide. For example, the terms "失眠 (insomnia)," "不寐 (sleeplessness)," and "辗转反侧 (nocturnal restlessness)" all convey the same concept and describe the symptom of "睡眠欠佳 (sleep disturbance)." However, only "失眠 (insomnia)" has a corresponding disease code in the *International Statistical Classification of Diseases, 10th Revision* (*ICD-10*) codes, so we set "失眠 (insomnia)" as the standard term. Under our fine-grained semantic classification rules, "失眠 (insomnia)" is categorized as the Symp_S entity. By aggregating such standard terms with the Symp_S label, we constructed the Symp_S standard knowledge base. In this study, symptom-related entities are categorized into 12 distinct types, necessitating the construction of 12 separate standard knowledge bases. Due to the limitations of the article's length, only the sources for the Symp_S and Pos_Pri standard terms are introduced.

We first describe the Symp_S standard knowledge base. In the field of medical information, the *ICD-10*, developed by the World Health Organization (WHO), has been widely used as a diagnostic coding and disease classification system in health records, information exchange, research, quality measurement, and payment applications [37]. We collected 136 Symp_S standard terms using *ICD-10* codes as the reference.

We then describe the Pos_Pri standard knowledge base. SNOMED CT is a comprehensive international clinical terminology system that contains more than 350,000 concepts and is organized into 19 top-level hierarchies, including body structure and clinical findings [38,39]. Because Pos_Pri entities are used to describe the primary site of a symptom, we constructed the Pos_Pri standard knowledge base by referring to the body structure hierarchy of SNOMED CT. The Pos_Pri standard knowledge base currently contains 106 commonly used anatomical terms.

Finally, we describe how the standard knowledge bases for the infrequent entity types were constructed. Entity types such as Time, Cond, Freq, and Trend are used less frequently in medical records. After searching for *ICD-10*, SNOMED CT, and the relevant literature on entity normalization, we were unable to find a dataset of these uncommon types of entities. To construct standard knowledge bases for these infrequent entities, we counted the number of occurrences of each extraction result after recognizing entities in 500 EMRs. Entities that occurred more than 10 times were designated as standard terms, after which domain experts manually removed duplicates with identical semantics.

Introduction to the Method of Constructing Positive and Negative Samples

We begin with the construction of positive samples. First, an NER model was used to extract all symptom entities contained in the sentence. Then, based on the label types of the symptom entities, experts manually identified the standard term with the same concept from the corresponding standard knowledge base. For instance, in the sentence "我的脑袋最近几天针扎样痛 (My head has had lancinating pain for several days)," the raw term "脑袋 (head)" was identified as a Pos_Pri entity. Therefore, to construct a positive sample about the entity "脑袋 (head)," we needed to find a synonym for "脑袋 (head)" in the Pos_Pri standard knowledge base. A detailed list of raw terms and their corresponding standard terms used for positive sample construction is available in Multimedia Appendix 1.

We then turn to the construction of negative samples. Negative samples were generated in batches by randomly combining the raw terms with unmatched standard terms of the same type. Take the sentence "我的胸口有刀割样痛 (I have a lancinating pain in my chest)" as an example. According to the result of named entity recognition, "刀割样痛 (lancinating pain)" was categorized as a State entity. The matched standard term in the State standard knowledge base was "绞痛 (colic)," while "酸痛 (soreness)" and "刺痛 (stabbing pain)" were unmatched standard terms in the same standard knowledge base. By combining "刀割样痛 (lancinating pain)" with "酸痛 (soreness)" and "刺痛 (stabbing pain)," respectively, negative samples were constructed for training the binary classification model.

The manual pairing required for positive samples (described above) contrasts with the batch generation possible for negative samples, which can lead to an imbalanced dataset. To address this potential imbalance, we used 5 data augmentation methods specifically to expand the positive sample set. The operational details and examples of these methods are provided in Multimedia Appendix 2.

We now present concrete examples of the constructed samples. We constructed a total of 3350 positive and negative samples. The ratio of positive to negative samples was 1:1. Examples are shown in Table 2, where label 1 represents that the raw term has the same concept as the standard term. Label 0 represents that the concepts of the 2 terms are different.

Table 2. Examples of positive and negative samples for the entity alignment model.

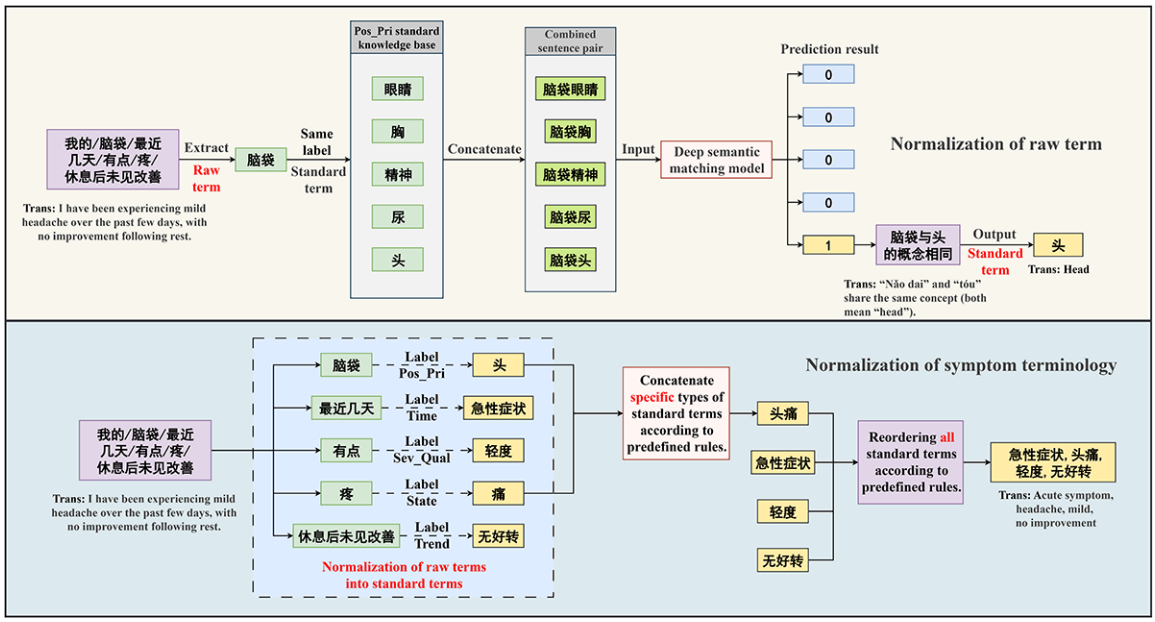
Raw term	Standard term	Label
Nǎo dai (脑袋)	Head	1
Pain tolerable, not affecting daily life or sleep quality, able to work	Mild [40]	1
Severe pain, unbearable, requiring analgesics, sleep quality impaired, but still able to continue working	Moderate [40]	1

Raw term	Standard term	Label
Severe pain, unbearable, requiring analgesics, sleep severely disturbed, unable to work, requiring hospitalization for systematic treatment	Severe [40]	1
24-hour urine output ≥3000 mL [41]	Polyuria	1
Hourly urine output ≤17 mL [41]	Oliguria	1
24-hour urine output ≤100 mL [42]	Anuria	1
Bowel movements fewer than 3 times per week [43]	Constipation	1
Bowel movements more than 3 times per day [44]	Diarrhea	1
Within the past few days	Acute symptoms	1
Occurring after climbing stairs	Exertion-induced	1
Lancinating pain	Colicky pain	1
Lancinating pain	Numbness	0
Lancinating pain	Aching pain	0

Methods of Normalizing Symptom Terminology

After inputting a paragraph of symptom terminology, the NER model was first used to extract raw terms. Each extracted raw term was paired with every standard term sharing the same fine-grained semantic label. Restricting candidate standard terms to the same semantic label reduced unnecessary comparisons across clinically incompatible entity types and made the alignment process more interpretable. This strategy is also consistent with the general formulation of biomedical entity linking, in which an entity mention is mapped to a standard concept or term in a predefined knowledge base [35]. Then, the binary classification model was used to determine whether the sentence pairs, obtained by concatenating the raw term and the standard term, represented a semantic match (label 1) or a nonmatch (label 0). Consistent with evidence that surface-form overlap is insufficient for resolving semantically ambiguous biomedical mentions [36], the alignment model was designed to judge semantic equivalence between raw terms and candidate standard terms rather than relying only on surface-level string similarity. A prediction of 1 signified semantic equivalence, prompting the system to output the corresponding standard term. If all predictions were 0, the raw term (eg, a Sev_Quant entity such as “about 5 kg”) was retained as the standard term because no matching result was found in the knowledge base. After normalizing all raw terms, standard terms were concatenated and reordered according to predefined rules to generate the final standardized symptom description. The normalization methods of raw terms and symptom terminology are shown in Figure 3.

Figure 3. Normalization methods for raw terms and symptom terminology.



Concatenation Rules for Standard Terms

To extract and reconstruct comprehensive symptoms from complex expressions with overlapping, discontinuous, and nested entities, after the deep semantic matching model normalized each extracted raw term to a standard term, we concatenated 6 categories of standard terms, including Neg, Symp_S, Pos_SCP, Pos_Pri, Pos_Sub, and State. This rule-guided reconstruction step was necessary because multi-implication TCM symptom expressions may contain

more than one clinically meaningful concept in a single raw phrase. Unlike conventional entity-linking settings that often assume a fixed mention span and map it to one knowledge-base concept [35], our task required both fine-grained decomposition of symptom components and structured recombination of multiple normalized components. After studying the raw medical records, we summarized the concatenation rules for standard terms into the following 5 types. The slash symbol "/" served as a delimiter between words.

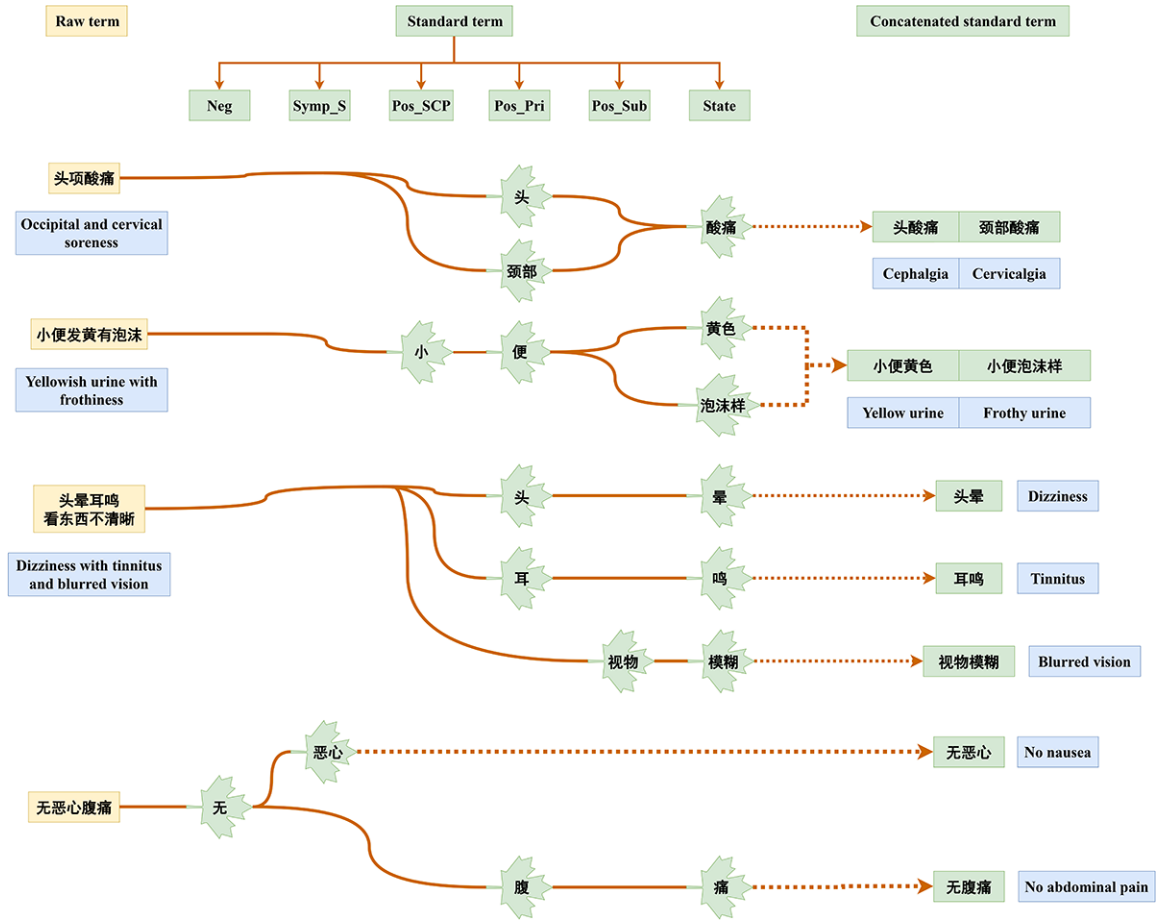
1. <Space$_1$+Space$_2$+...+State>: This structure indicates concurrent symptoms in multiple anatomical locations (eg, 头/项/酸痛 [occipitocervical aching and headache]).
2. <Space+State$_1$+State$_2$+...>: This structure denotes the concurrent presence of 2 or more distinct symptoms within a single anatomical location, physiological product, or mental state (eg, 小/便/发黄/有泡沫 [dark-colored urine with foam formation]).
3. <Space$_1$+State$_1$+Space$_2$+State$_2$>: Within such structurally complex expressions, multiple Space entities coexist alongside multiple State entities. Comprehensive restoration of the symptoms requires a 2-by-2 combination of the Space and the State entities (eg, 头/晕/耳/鸣 [vertigo with concomitant tinnitus]).
4. <Space+Sev_Qual+State> and <Space+Freq+State>: In these 2 types of structurally complex expressions, between a specific anatomy (or physiological product) and a specific symptom, there are adjectives or adverbs to further describe the severity or frequency of the symptom in greater detail (eg, 头/有点/痛 [a mild headache], 大/便/经常/困难 [chronic defecatory difficulty]). To reconstruct the described symptom, other types of entities that exist between Space and State must be temporarily ignored, prioritizing the combination of the Space and State entities.
5. <Space+Neg+State$_1$+State$_2$+...>, <Neg+Symp_S$_1$+Symp_S$_2$+...>, <Neg+Symp_S+Space+State>, <Neg+Space$_1$+State$_1$+Space$_2$+State$_2$>: These complex structures are usually used to enumerate 2 or more negative symptoms (eg, 关节/无/红肿/热痛 [absence of joint erythema, swelling, hotness, or tenderness], 无/发热/恶寒 [absence of fever and chills]).

It is worth noting that the Space entity can be categorized into 2 distinct types: physical space and conceptual space. Physical space refers to entities that can be directly seen and touched, such as various parts of the human body and different physiological products, including urine, feces, sweat, and menstruation. Conceptual space, on the other hand, is an intangible, abstract existence that usually includes organ functions, numerous sensations, and mental activities. Thus, the Space entity that combines with the State entity can be further subdivided into the following five types: (1) an independent Pos_Pri (eg, 头 [head]); (2) Pos_SCP+Pos_Pri (eg, 小/便 [urine]); (3) an independent Pos_Sub (eg, 视物 [vision] and 皮肤 [skin]); (4) Pos_Pri+Pos_Sub (eg, 腕/关节 [wrist joint]); and (5) Pos_SCP+Pos_Pri+Pos_Sub (eg, 小/便/量 [urine volume]).

From the above combination structures, it can be seen that (1) the State entity cannot exist alone and must be concatenated with the previous Space entity to show clinical diagnostic value; (2) the Space entity that can be concatenated with the State entity is not limited to specific anatomy but also includes secretions such as urine and feces, as well as intangible physiological functions and mental states. The Space entity is broader than the Pos_Pri entity; (3) positive symptoms consist of 2 conceptual elements: Space and State, while negative symptoms are composed of 3 conceptual elements: Neg, Space, and State. Before outputting the concatenated results, it is necessary to check whether any negative symptoms have not yet been concatenated to avoid incorrectly converting them into positive ones.

Based on the above 3 core principles, we constructed a trie tree to concatenate the extracted standard terms: each branch (path) represents a complete symptom standard term. The leaves (start points of paths) are State or Symp_S entities. The roots (end points of paths) are Space or Neg entities, and each node on the branches (paths) has only one standard term. By collecting all branches (symptom standard terms), all symptoms in the complex expression can be extracted and comprehensively reconstructed. The trie tree concatenation method of standard terms is shown in Figure 4.

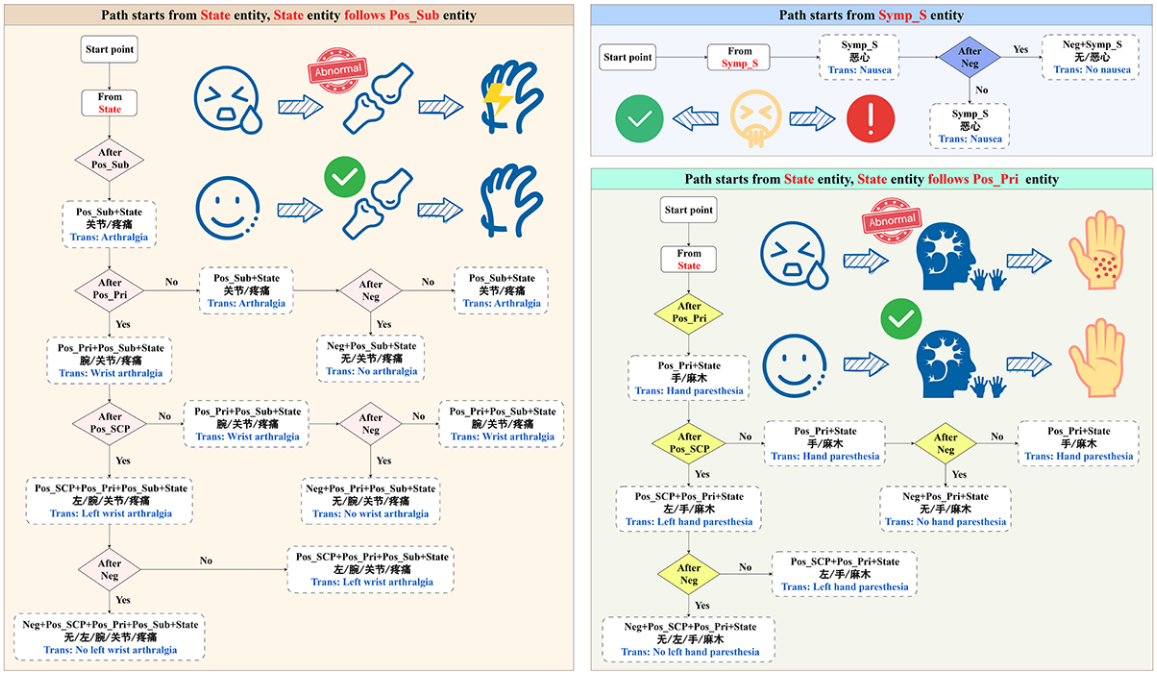
Figure 4. Trie tree concatenation method of standard terms.



When the path starts from the State entity, it connects with the Pos_Pri or Pos_Sub node. If the node connected to the State entity is Pos_Sub, continue to determine whether it needs to connect with the Pos_Pri node, Pos_SCP node, and Neg node. If no additional nodes are available for concatenation, end the current path, and the symptom standard term becomes the output. When the node connected to the State node is Pos_Pri, continue to determine whether it needs to connect with the Pos_SCP node and Neg node.

When the path starts from a Symp_S entity and no Neg node can be used for concatenation in front of the Symp_S entity, the path can be ended directly, and the symptom standard term can be output. The specific method of path generation is shown in Figure 5.

Figure 5. Flowchart of path generation.



Sorting Rules of Standard Terms

Reordering standard terms yielded structured, coherent, and clinically meaningful outputs. Analysis of large-scale EMRs revealed that physicians typically follow a specific logical sequence when documenting: (1) onset time and precipitating factors; (2) symptom manifestations, severity, and frequency; and (3) disease progression. Therefore, after concatenating standard terms through the trie tree method, the final normalized output follows this order: Time entity, Cond entity, concatenated standard term, Sev_Qual entity, Sev_Quant entity, Freq entity, and Trend entity.

Experimental Evaluation Indicators

The framework was evaluated at 2 stages, named entity recognition and named entity normalization, each with its own set of metrics. For the task of NER, model performance is commonly evaluated using precision, recall, and F_1-score. The calculation of these metrics is based on the following classifications of the prediction results:

1. True positive (TP): represents the number of correctly identified entities.
2. True negative (TN): represents the number of correctly identified nonentities.
3. False positive (FP): represents the number of instances incorrectly identified as entities.
4. False negative (FN): represents the number of entities that the model failed to identify.

The specific formulas and their interpretations are shown in Table 3. As the F_1-score provides a more comprehensive evaluation by considering both precision and recall, it was selected as the primary performance evaluation metric for the NER model in this study.

Table 3. Formulas and interpretations of evaluation metrics for named entity recognition (NER).

Evaluation metric	Equation	Interpretation
Precision	$P=\frac{TP}{TP+FP}$	This metric, also known as positive predictive value, represents the proportion of true entities among all results identified as entities
Recall	$R=\frac{TP}{TP+FN}$	This metric, also known as sensitivity, indicates the proportion of all actual entities that are successfully identified by the model
F_1-score	$F_1=\frac{2PR}{P+R}$	This metric is the harmonic mean of precision and recall

We then describe the metrics used for the normalization stage. For the NEN task, accuracy, precision, recall, and F_1-score were used to evaluate model performance. Accuracy was calculated at the instance level using exact matching as the criterion. A prediction was considered correct only when the complete set of predicted standard terms is identical to the complete set of reference standard terms. Precision, recall, and F_1-score were calculated at the standard-term level to further assess the ability of each model to retrieve individual standard terms. The formulas of these evaluation metrics are shown in Table 4.

Table 4. Formulas of evaluation metrics for named entity normalization (NEN).

Evaluation metric	Equation
Accuracy	$\text{Acc} = \frac{\text{Number of instances with completely correct prediction results}}{\text{Total number of test instances}}$
Precision	$\text{P} = \frac{\text{Number of correctly predicted standard terms}}{\text{Total number of standard terms output by the model}}$
Recall	$\text{R} = \frac{\text{Number of correctly predicted standard terms}}{\text{Total number of reference standard terms}}$
F_1-score	$F_1 = \frac{2\text{PR}}{\text{P+R}}$

To verify the performance differences between the proposed STC-NF model and the baseline models, the held-out test set generated from the 8:2 split of the 500 collected EMRs, as described in the "Introduction to Dataset Sources and Fine-Grained Semantic Classification" subsection, was further divided into 3 NEN test sets. First, 25 records were randomly selected, and the single-implication terms in these records were used to construct the single-implication test set. Second, another 25 records were randomly selected, and the multi-implication terms contained in these records were used to construct the multi-implication test set. Finally, the remaining 50 records were used to construct the mixed test set, which contained both single- and multi-implication terms. Therefore, the 3 test datasets were constructed from nonoverlapping EMRs, and the multi-implication terms in the mixed test set did not overlap with those in the multi-implication test set.

Experimental Hyperparameters of the NEN Model

We used BERT-base-Chinese as the base model, which had a hidden layer dimensionality of 768. Because the NEN module formulated entity alignment as a sentence-pair semantic matching task between a raw term and a candidate standard term, the hyperparameter settings were determined by considering previous studies of medical concept normalization and biomedical entity normalization, as well as the short phrase-level characteristics of TCM symptom-term pairs. Recent Chinese medical concept normalization and rare disease concept normalization studies have shown that deep language models can effectively map nonstandard medical expressions to standardized medical concepts or vocabulary identifiers [45,46]. In a recent NER/NEN corpus study, transformer fine-tuning hyperparameters were selected on a development set, and a maximum sequence length of 128 was used for short biomedical entity mentions [47]. In addition, recent biomedical BERT fine-tuning experiments commonly evaluated learning rates of 1×10^{-5}, 3×10^{-5}, and 5×10^{-5}, sequence lengths of 128, 256, and 512, batch sizes of 16 and 32, and dropout of 0.1 [48]. Therefore, we set the learning rate to 5×10^{-5}, which falls within the commonly evaluated BERT fine-tuning range and allows efficient adaptation on a relatively small entity-alignment dataset. A dropout rate of 0.1 was used to reduce the risk of overfitting during supervised sentence-pair classification. The maximum sequence length was set to 128 because both the raw TCM symptom terms and candidate standard terms were short, phrase-level inputs; this setting was sufficient to cover the term pairs while reducing unnecessary computation. The training batch size was set to 30, close to the commonly used mini-batch range in biomedical BERT fine-tuning, while remaining compatible with the available GPU memory. The Adam optimizer was used for gradient-based parameter updates during BERT fine-tuning.

Ethical Considerations

This study involving humans was approved by the Second Traditional Chinese Medicine Hospital of Guangdong Province, affiliated with the Guangzhou University of Chinese Medicine. The ethics approval number was Y202407-004-01. The study was conducted according to local legal and institutional requirements and adhered to the Declaration of Helsinki. This study was retrospective, and informed consent was waived.

Results

Named Entity Recognition Results

In this study, symptom terms were categorized into 12 different types of entities based on fine-grained semantic classification. Three NER models specialized for extracting symptom entities were trained according to customized annotation rules. The extraction performance of each model is shown in Table 5.

Table 5. Comparison of symptom entity extraction performance across models.

Entity type	BiLSTM[a]-CRF[b] F_1[c]-score ratio, (%)	BERT[d]-CRF F_1-score ratio, (%)	BERT-BiLSTM-CRF F_1-score ratio, (%)
Independent symptom (Symp_S)	1100/1186 (92.7)	1082/1186 (91.2)	1106/1192 (92.8)
Negative occurrence (Neg)	1112/1160 (95.9)	1088/1148 (94.8)	1102/1153 (95.6)
Scope space (Pos_SCP)	880/915 (96.2)	874/910 (96.0)	882/916 (96.3)
Primary space (Pos_Pri)	4324/4711 (91.8)	4316/4702 (91.8)	4340/4707 (92.2)
Secondary space (Pos_Sub)	348/370 (94.1)	344/365 (94.2)	348/368 (94.6)
Time of occurrence (Time)	816/888 (91.9)	796/880 (90.5)	804/884 (91.0)
Occurrence condition (Cond)	332/376 (88.3)	342/386 (88.6)	342/383 (89.3)
Frequency of occurrence (Freq)	416/446 (93.3)	406/441 (92.1)	414/443 (93.5)
Qualitative severity (Sev_Qual)	450/490 (91.8)	450/488 (92.2)	458/490 (93.5)
Quantitative severity (Sev_Quant)	90/103 (87.4)	92/105 (87.6)	92/104 (88.5)
Occurrence state (State)	4530/5124 (88.4)	4534/5112 (88.7)	4586/5141 (89.2)
Occurrence trend (Trend)	170/207 (82.1)	174/204 (85.3)	178/208 (85.6)

[a]BiLSTM: bidirectional long short-term memory.
[b]CRF: conditional random field.
[c]In every model column, each cell reports the F_1-score as 2TP/(2TP+FP+FN) with the resulting percentage in parentheses (TP, FP, and FN as defined in the Experimental Evaluation Indicators subsection); this ratio is the F_1-score calculation and does not represent the count of correctly recognized entities divided by the total number of entities.
[d]BERT: Bidirectional Encoder Representations from Transformers.

As shown in Table 5, the 3 models achieved comparable overall performance, although the best-performing model varied across entity types. BERT-BiLSTM-CRF obtained the highest or tied-highest F_1-score for most entity types, achieving 4340 out of 4707 (92.2%) for Pos_Pri and 4586 out of 5141 (89.2%) for State among the high-frequency structural entities, as well as 1106 out of 1192 (92.8%) for Symp_S, 882 out of 916 (96.3%) for Pos_SCP, and 458 out of 490 (93.5%) for Sev_Qual. BiLSTM-CRF achieved the highest F_1-scores for the Time and Neg entities, with F_1-scores of 816 out of 888 (91.9%) and 1112 out of 1160 (95.9%), respectively, and remained within approximately one percentage point of the best-performing model on the high-frequency structural entities. The advantages of the BERT-based models over BiLSTM-CRF were concentrated on low-frequency modifier entities, such as Sev_Quant, Sev_Qual, and Trend, where the best-performing BERT-based model exceeded BiLSTM-CRF by 1.1, 1.7, and 3.5 percentage points, respectively. Complete per-entity precision, recall, and F_1-scores for all three models are provided in Multimedia Appendix 3.

Entity Alignment Results

We used 3 statistical models (TF-IDF, Med, and BM25), the multiclassification model PLM-*ICD*, the generative sequence-generation model, and the binary classification model MTCG as baseline models for comparison with our proposed STC-NF model. The experimental results of different normalization models on 3 test sets are shown in Table 6.

Table 6. Results of different normalization models on 3 test sets.

Model name and test set	Accuracy[a], n/N (%)	Precision, n/N (%)	Recall, n/N (%)	F_1-score, ratio (%)
TF-IDF[b]				
Single-implication	91/197 (46.2)	91/186 (48.9)	91/197 (46.2)	182/383 (47.5)
Multi-implication	0/369 (0.0)	83/328 (25.3)	83/1175 (7.1)	166/1503 (11.0)
Mixed	179/1124 (15.9)	344/1068 (32.2)	344/2225 (15.5)	688/3293 (20.9)
Med				
Single-implication	97/197 (49.2)	97/188 (51.6)	97/197 (49.2)	194/385 (50.4)
Multi-implication	0/369 (0.0)	88/331 (26.6)	88/1175 (7.5)	176/1506 (11.7)
Mixed	192/1124 (17.1)	366/1070 (34.2)	366/2225 (16.4)	732/3295 (22.2)
BM25				
Single-implication	106/197 (53.8)	106/191 (55.5)	106/197 (53.8)	212/388 (54.6)
Multi-implication	0/369 (0.0)	92/337 (27.3)	92/1175 (7.8)	184/1512 (12.2)
Mixed	210/1124 (18.7)	391/1082 (36.1)	391/2225 (17.6)	782/3307 (23.6)
PLM-*ICD*[c]				
Single-implication	168/197 (85.3)	168/184 (91.3)	168/197 (85.3)	336/381 (88.2)

Model name and test set	Accuracy[a], n/N (%)	Precision, n/N (%)	Recall, n/N (%)	F_1-score, ratio (%)
Multi-implication	0/369 (0.0)	97/343 (28.3)	97/1175 (8.3)	194/1518 (12.8)
Mixed	326/1124 (29.0)	522/1102 (47.4)	522/2225 (23.5)	1044/3327 (31.4)
Sequence generation				
Single-implication	157/197 (79.7)	157/180 (87.2)	157/197 (79.7)	314/377 (83.3)
Multi-implication	189/369 (51.2)	593/1134 (52.3)	593/1175 (50.5)	1186/2309 (51.4)
Mixed	706/1124 (62.8)	1245/2142 (58.1)	1245/2225 (56.0)	2490/4367 (57.0)
MTCG[d]				
Single-implication	173/197 (87.8)	173/188 (92.0)	173/197 (87.8)	346/385 (89.9)
Multi-implication	223/369 (60.4)	698/1137 (61.4)	698/1175 (59.4)	1396/2312 (60.4)
Mixed	803/1124 (71.4)	1443/2156 (66.9)	1443/2225 (64.9)	2886/4381 (65.9)
STC-NF[e]				
Single-implication	180/197 (91.4)	180/192 (93.8)	180/197 (91.4)	360/389 (92.5)
Multi-implication	311/369 (84.3)	979/1141 (85.8)	979/1175 (83.3)	1958/2316 (84.5)
Mixed	990/1124 (88.1)	1931/2160 (89.4)	1931/2225 (86.8)	3862/4385 (88.1)

[a]Accuracy, precision, and recall are reported as n/N (%), whereas F_1-score is reported as a ratio (%) calculated as 2TP/(2TP+FP+FN). Accuracy is calculated at the instance level using exact matching; precision, recall, and F_1-score are calculated at the standard-term level. See Table 4 for the corresponding formulas.
[b]TF-IDF: term frequency-inverse document frequency.
[c]PLM-*ICD*: pretrained language model framework for *International Classification of Diseases* coding.
[d]MTCG: multi-task candidate generator.
[e]STC-NF: split-then-concatenate normalization framework.

As shown in Table 6, STC-NF achieved the best overall performance across the 3 test sets. On the single-implication test set, STC-NF achieved an accuracy of 180 out of 197 (91.4%), a precision of 180 out of 192 (93.8%), a recall of 180 out of 197 (91.4%), and an F_1-score of 360 out of 389 (92.5%). Compared with PLM-*ICD* and MTCG, which were the stronger baseline models for single-implication terms, STC-NF improved accuracy by 6.1 and 3.6 percentage points, respectively.

For the multi-implication test set, the accuracy of TF-IDF, Med, BM25, and PLM-*ICD* was 0 out of 369 (0.0%). This result was attributable to the single-output design of these models, which prevented them from producing a complete set of standard terms for multi-implication symptom terms. However, their term-level precision, recall, and F_1-score were still calculated to reflect partial standard-term retrieval performance. In contrast, STC-NF achieved an accuracy of 311 out of 369 (84.3%), a precision of 979 out of 1141 (85.8%), a recall of 979 out of 1175 (83.3%), and an F_1-score of 1958 out of 2316 (84.5%). Compared with sequence generation and MTCG, STC-NF improved multi-implication accuracy by 33.1 and 23.9 percentage points, respectively.

On the mixed test set, STC-NF achieved an accuracy of 990 out of 1124 (88.1%), a precision of 1931 out of 2160 (89.4%), a recall of 1931 out of 2225 (86.8%), and an F_1-score of 3862 out of 4385 (88.1%). These results were higher than those of all baseline models. Compared with MTCG, the best-performing baseline model on the mixed test set, STC-NF improved accuracy by 16.7 percentage points and F_1-score by 22.2 percentage points. These findings indicate that the split-then-concatenate mechanism improves not only complete instance-level normalization accuracy but also standard-term-level retrieval performance, particularly for complex multi-implication symptom terms.

Structured Normalization Results

After identifying the standard terms with the same concept as each raw term through a BERT-based binary classification model, this study concatenated and reordered standard terms with specific semantic labels according to predefined rules and output structured text to complete the normalization of TCM symptom terminology. To improve the readability of the main text and avoid an overly crowded bilingual table, Tables 7 and 8 present the English versions of the structured normalization results. The corresponding original Chinese raw terms, segmentation results, and structured normalization outputs are available in Multimedia Appendix 4.

Table 7. Translated examples of raw symptom expressions and their normalized symptom terms.

Number	Raw term	Concatenated standard term
1	Occipitocervical aching and headache have been reported in recent days, without significant improvement following rest.	Headache and neck pain
2	Mild exertional chest tightness and dyspnea on stair climbing have persisted for several weeks.	Chest tightness and shortness of breath
3	An unintentional weight loss of approximately 5 kg has occurred over the past 6 months.	Weight loss

Number	Raw term	Concatenated standard term
4	Severe persistent odynophagia has been noted without associated chills or fever.	Sore throat without chills or fever

Table 8. Structured semantic attributes corresponding to the examples in Table 7.

Number	Time	Cond	Sev_Qual	Sev_Quant	Freq	Trend
1	Acute symptom	—[a]	—	—	—	Not improving
2	Subacute symptom	Exertion-induced	Mild	—	—	—
3	Chronic symptom	—	—	Approximately 5 kg	—	—
4	—	—	Severe	—	Persistent symptom	—

[a]Not applicable.

Discussion

Performance Comparison of Named Entity Recognition Models

In the experiments of recognizing symptom entities, BiLSTM-CRF achieved the highest F_1-scores among the 3 models for the Time and Neg entities, at 91.9% and 95.9%, respectively. For several entity types, including Pos_Pri, Pos_Sub, and State, the differences in F_1-scores among BiLSTM-CRF, BERT-CRF, and BERT-BiLSTM-CRF were small (within about one percentage point), with BiLSTM-CRF reaching an F_1-score of 91.8% for the Pos_Pri entity. Importantly, although the BERT-based models obtained higher F_1-scores for most entity types, their largest gains over BiLSTM-CRF were concentrated on low-frequency modifier entities, such as Sev_Quant, Sev_Qual, and Trend. In contrast, for the high-frequency structural entities that dominate downstream symptom reconstruction—most notably Pos_Pri and State—BiLSTM-CRF was essentially on par with the pretrained models, trailing by only about 0.4 and 0.8 percentage points, respectively. This narrow gap can be attributed to the fine-grained semantic classification scheme adopted in this study. By constraining each category to short, structurally regular spans with clearly delimited boundaries, the annotation rules reduce intracategory description diversity and lower the decision dimensionality of the recognition model so that a lightweight sequence model is already able to capture the local dependencies required for accurate boundary and type prediction. Consequently, the marginal benefit obtainable from large-scale pretraining is limited under the present annotation scheme. Xue et al [49] and Li et al [50] proposed the porous lattice transformer encoder (PLTE) and flat-lattice transformer (FLAT), respectively, for Chinese NER, showing that incorporating external lexicon information through character-word lattices can effectively exploit word-boundary cues and improve Chinese NER performance. These studies suggest that lexicon-enhanced transformer models are valuable when word-boundary ambiguity and external word information are major factors affecting entity recognition. However, the objective of the NER stage in STC-NF was not to pursue the highest possible benchmark score but to provide stable fine-grained semantic units for subsequent label-constrained entity alignment and rule-guided reconstruction. Considering the recognition performance reported above, its lower model complexity (specifically, substantially fewer parameters and no dependence on large-scale pretrained weights), and the downstream requirement for stable entity decomposition, BiLSTM-CRF was selected as the NER component of STC-NF. Although the BERT-based models achieved marginally higher F_1-scores on most entity types in Table 5, the largest gains occurred on low-frequency modifier entities, which are peripheral to symptom reconstruction and can be further rectified by the downstream concatenation and sorting rules. Therefore, this advantage does not warrant their substantially higher computational cost in the present 2-stage pipeline. More advanced lexicon-enhanced transformer architectures, such as PLTE and FLAT, remain promising alternatives for future evaluation on larger and more heterogeneous TCM clinical corpora.

Merits and Limitations of Baseline Models

Although statistical models have the advantage of fast processing, they are not competitive with deep learning models in semantic understanding. Clinical term normalization studies have shown that nonstandard clinical expressions often differ from standardized terminology because of synonyms, stylistic variation, morphological variation, abbreviations, and local writing conventions, making surface-form matching insufficient for robust normalization [51]. Similarly, recent biomedical NER and NEN systems have combined neural NER with neural or hybrid normalization modules to overcome the inability of dictionary- or rule-only methods to cover complex mention variations [52]. Therefore, the relatively weak performance of TF-IDF, Med, and BM25 in this study was mainly attributable to their dependence on character-level or lexical similarity rather than contextual semantic matching.

The multiclassification model PLM-*ICD* first segments the raw text into sentences. After embedding each sentence and all *ICD* codes into dense vectors, it selects the *ICD* embedding with the highest cosine similarity (computed using attention-weighted BERT or BioBERT models pretrained on corpora, such as Medical Information Mart for Intensive Care III [MIMIC-III]), to align the raw term. PLM-*ICD* performed well for single-implication terms. However, because of its multiclassification design, the model could output only one prediction for each input sentence, regardless of how many

symptoms were included. Therefore, PLM-*ICD* could not handle multi-implication terms such as "腰酸背痛 (lumbago and back pain with soreness)," "尿频量多 (polyuria with increased urinary frequency)," and so on.

Generative model sequence generation first uses a character-level text generation model built upon transformer to convert each character in the raw term step by step. Once multiple candidate entity sets (sentences) of standard terms are generated, a pretrained BERT model calculates the semantic similarity between the input text and each set. Then, the top-n generated sets are chosen as the final normalization results after reordering. Compared to PLM-*ICD*, this model demonstrates the capability to handle multi-implication terms. However, the single-implication accuracy is significantly lower, as splitting the raw term word by word sometimes destroys the structural integrity of the entity. For example, if "恶心 (nausea)" is split into "恶 (hate)" and "心 (heart)" and then converted and concatenated together, the final result will be the wrong prediction of "厌恶心脏 (loathe heart)."

The binary classification model MTCG significantly reduces the number of semantic comparisons through its retrieval strategy, resulting in higher entity alignment efficiency than the multiclassification approach. In addition, it achieved higher accuracy in dealing with single-implication terms such as "恶心 (nausea)," "发热 (fever)," "咳 (cough)," etc, which contain only one symptom, because it does not split the raw term and retains the structural integrity of the entity, compared to the sequence-generation approach. A key limitation of this model is its misclassification of compound symptoms that have unclear entity boundaries. The advantage of STC-NF over MTCG was most evident for multi-implication terms. Unlike MTCG, STC-NF explicitly decomposes multi-implication expressions before semantic matching and reconstructs the normalized outputs after alignment. This additional split-and-recombine process helps preserve multiple symptom concepts within a single raw phrase, which cannot always be represented by a single retrieved standard term. Taking "尿频量多 (polyuria with increased urinary frequency)" as an example, since "量多 (high volume)" does not immediately follow "尿 (urine)," MTCG recognized only the raw term "尿频 (frequent urination)" but omitted another raw term "尿量多 (excessive urination)" at the end. Therefore, although MTCG exhibits remarkable generalization performance and excellent accuracy on the mixed test set, a significant gap remains between MTCG and our model when handling multi-implication terms.

Mechanistic Explanation for the Superior Performance of STC-NF

Overall, the superior performance of STC-NF can be explained by the complementarity of its 3 components. First, fine-grained semantic extraction decomposes complex symptom expressions into clinically meaningful semantic units, thereby reducing the boundary ambiguity caused by overlapping, discontinuous, and nested entities. This is consistent with recent medical NER evidence showing that entity phrase length and the number of words in an entity phrase can substantially influence recognition performance [53], and with systematic evidence that discontinuous clinical entities remain challenging for traditional NER methods [54]. Second, the BERT-based binary classification module preserves semantic matching at the concept-alignment stage instead of relying only on surface similarity. Third, the trie tree–based concatenation rules explicitly reconstruct multi-implication symptom terms after semantic alignment. This design is consistent with 2-stage clinical text workflows that first extract information from free-text notes and then harmonize extracted mentions into standardized database concepts [55], as well as recent hybrid neural-symbolic clinical NLP systems that integrate neural recognition with structured vocabularies and symbolic reasoning to transform unstructured clinical notes into standardized terms [56]. Therefore, STC-NF improved not only complete instance-level normalization accuracy but also standard-term-level retrieval performance, especially for multi-implication symptom terms with complex internal structures.

Limitations and Future Work

This study has several limitations that suggest avenues for future research. First, the NEN module used BERT-base-Chinese as the base semantic-matching model, without evaluating more recent clinical or biomedical language–modeling strategies. Indirect evidence from our previous work on the upstream NER stage suggests that the choice of pretrained backbone is important for fine-grained TCM symptom modeling. In that study, we evaluated several external NER backbone models on fine-grained TCM symptom corpora, including BERT-base-Chinese+CRF, BERT-base-Chinese+BiLSTM+CRF, Chinese-RoBERTa-wwm-ext+CRF, and Chinese-RoBERTa-wwm-ext+BiLSTM+CRF. The results showed that Chinese-RoBERTa-wwm-ext+BiLSTM+CRF achieved higher NER F_1-scores than BiLSTM+CRF on both GDTCM-500 and HwaMei-500, suggesting that more advanced pretrained language models may further improve symptom entity recognition [57]. However, these experiments were limited to the entity recognition stage. We have not yet extended these external backbone comparisons to the full NEN stage, because such validation would require reconstructing entity-alignment samples, rerunning the complete split-then-concatenate pipeline, and performing additional expert validation of normalized outputs. Therefore, future work will compare STC-NF with more advanced pretrained language models, retrieval-augmented models, and domain-specific clinical language models at both the NER and NEN stages to determine whether the proposed split-then-concatenate mechanism remains beneficial across different backbone architectures. Accordingly, these results should be interpreted as evidence for the effectiveness of the split-then-concatenate mechanism under a BERT-base-Chinese alignment setting, rather than as evidence that BERT-base-Chinese is the optimal backbone for the NEN module.

Second, the generalizability of the model may be constrained, as the training data were sourced entirely from a single clinical department (endocrinology). The framework's effectiveness has not yet been validated using multisource

data from departments such as surgery or pediatrics. This limitation is particularly important because model robustness depends not only on algorithmic structure but also on how heterogeneous information sources are represented and transferred across contexts. Beyond medical applications, the methodological idea of quantifying free-text content and integrating it with structured variables in decision-support pipelines has been demonstrated in other domains as well, for example, by Wang et al [58] in a consumer-review setting. More directly within health care, Hou et al [59] highlighted in a chronic disease prediction setting that multimodal electronic health record representation learning and domain adaptation can improve predictive generalizability under cross-domain distribution shifts. These studies suggest that future versions of STC-NF should not only expand the corpus to multiple clinical departments but also explore whether domain adaptation and multimodal representation learning can improve the robustness of TCM symptom terminology normalization across heterogeneous EMR sources.

Finally, the construction of positive samples and standard knowledge bases still relied partly on manual expert curation, especially for infrequent entities. This process ensured semantic accuracy but limited scalability. Recent work on medical concept annotation has noted that supervised medical concept normalization often requires sufficient annotated data, whereas the availability of annotated clinical data is constrained by privacy, sensitivity, and the time required for expert annotation [60]. Similarly, clinical registry-oriented NLP reviews have identified manual extraction and annotation as labor- and resource-intensive steps that may hinder large-scale deployment [61]. To reduce this bottleneck, future work will explore semiautomatic candidate generation and weakly supervised concept annotation. In addition, the distinctive features of Chinese radicals may provide useful linguistic cues for TCM symptom terminology. For example, Chinese characters containing the radical "疒" are often related to State entities, such as "疼 (painful)," "瘫 (paralytic)," and "疯 (mad)," whereas characters containing the radical "月" are often found in Pos_Pri entities related to anatomy, such as "胸 (chest)," "腹 (abdomen)," and "肢 (limb)." Incorporating radical-aware representations into language models may reduce the candidate screening range and improve the efficiency of positive sample construction.

Conclusions

This study proposed the STC-NF, a 2-stage deep learning approach that uses fine-grained semantic classification to normalize TCM symptom terminology. The STC-NF model achieved an accuracy of 180 out of 197 (91.4%) on the single-implication test set, 311 out of 369 (84.3%) on the multi-implication test set, and 990 out of 1124 (88.1%) on the mixed test set. In addition, the corresponding F_1-scores reached 360 out of 389 (92.5%), 1958 out of 2316 (84.5%), and 3862 out of 4385 (88.1%), respectively. These results demonstrate that the split-then-concatenate mechanism substantially improves both instance-level normalization accuracy and standard-term-level retrieval performance for TCM symptom terminology normalization.

Beyond its quantitative performance, the proposed framework facilitates the automatic extraction and structuring of key clinical information from unstructured EMRs, thereby enabling the development of advanced applications. In clinical practice, these structured data can enable intelligent analyses to support clinicians in diagnosis and treatment. By converting diverse and ambiguous patient-reported symptoms into standardized terms, the framework provides a consistent data foundation for clinical decision support systems. This uniformity allows for large-scale analysis to identify symptom patterns predictive of specific conditions, potentially reducing diagnostic ambiguity and enhancing the precision of care.

Furthermore, this data structuring capability is instrumental for constructing high-quality biomedical knowledge bases and knowledge graphs. The normalization process transforms raw textual descriptions into discrete, standardized entities, a prerequisite for building robust and interoperable knowledge representations. Such knowledge graphs can map complex relationships among symptoms, diseases, and treatments, potentially facilitating the development of sophisticated tools for intelligent health care and personalized medicine. However, practical deployment of this framework still requires sufficiently representative annotated EMR data, continuously updated standard knowledge bases, and further validation across heterogeneous clinical departments.

Finally, this work has significant implications for health care administration, particularly for medical insurance cost control and diagnosis-related group (DRG) assignment. Hospital reimbursement through DRGs depends in part on accurate diagnostic coding derived from clinical evidence, including patient symptoms. The heterogeneity in raw symptom descriptions can lead to inconsistent coding and inaccurate DRG assignments. By standardizing symptom terminology, our framework produces uniform, machine-readable data that enhance diagnostic reliability. This improvement fosters more consistent and accurate coding, which is fundamental to a precise DRG system. Consequently, our approach supports a more scientific and equitable assessment of hospital management and service quality, ensuring that reimbursement accurately reflects patient case complexity and contributing to the advancement of medical informatization.

Acknowledgments

The authors disclose that no generative AI was used in any portion of the manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (grant 2023YFC3503002), the 2026 "Unveiling and Commanding" Project of the School of Medical Information Engineering at Guangzhou University of Chinese Medicine (grant 2026-7), and the Platform Project of the Big Data Research Center for Traditional Chinese Medicine at Guangzhou University of Chinese Medicine (grant A1-2601-25-439-127Z102).

Data Availability

The datasets generated and analyzed during this study are not publicly available due to patient privacy considerations. However, deidentified data may be made available from the corresponding author upon reasonable request.

Authors' Contributions

J Yao: Writing - review & editing, Writing - original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. XG: Writing - review & editing, Software, Methodology, Data curation, Conceptualization. WL: Writing - review & editing, Software, Methodology. YG: Writing - review & editing, Data curation. SW: Writing - review & editing. CZ: Writing - review & editing, Data curation. HY: Methodology, Formal analysis. JT: Writing - review & editing. J Yi: Writing - review & editing, Resources, Methodology, Conceptualization. DC: Writing - review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Multimedia Appendix 1

Examples of the positive sample dataset.
[DOC File (Microsoft Word File), 3063 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Introduction to positive sample data augmentation methods.
[DOC File (Microsoft Word File), 3941 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Complete per-entity precision, recall, and F_1-score values for all models.
[DOC File (Microsoft Word File), 179 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Original Chinese examples of structured normalization results.
[DOC File (Microsoft Word File), 3060 KB-Multimedia Appendix 4]

References

1. Wang L, Tang K, Wang Y, Zhang P, Li S. Advancements in artificial intelligence-driven diagnostic models for traditional Chinese medicine. Am J Chin Med. 2025;53(3):647-673. [doi: 10.1142/S0192415X25500259] [Medline: 40374369]
2. Tang Z, Wu Q, Jia Y, et al. Research progress on knowledge discovery in traditional Chinese medicine medical records based on artificial intelligence. Guidelines Stand Chin Med. 2024;2(4):174-181. [doi: 10.1097/gscm.0000000000000035]
3. Huang L, Wang Q, Duan Q, et al. TCMSSD: a comprehensive database focused on syndrome standardization. Phytomedicine. Jun 2024;128:155486. [doi: 10.1016/j.phymed.2024.155486] [Medline: 38471316]
4. Song M, Ni F, Li J, Li X, Li K. Status quo, problem, and prospect for traditional Chinese medicine international terminology standards. Guidelines Stand Chin Med. 2025;3(1):1-7. [doi: 10.1097/gscm.0000000000000052]
5. Shu Z, Hua R, Yan D, et al. ISPO: an integrated ontology of symptom phenotypes for semantic integration of traditional Chinese medical data. Methods Inf Med. Dec 2024;63(5-06):164-175. [doi: 10.1055/a-2576-1847] [Medline: 40328309]
6. Jia Q, Zhang D, Yang S, et al. Traditional Chinese medicine symptom normalization approach leveraging hierarchical semantic information and text matching with attention mechanism. J Biomed Inform. Apr 2021;116:103718. [doi: 10.1016/j.jbi.2021.103718] [Medline: 33631381]
7. Zhan Y, Zhang D, Jia Q, Xu H, Xie Y. Automated methods for symptom normalization in traditional Chinese medicine records. In: Sun X, Zhang X, Xia Z, Bertino E, editors. Advances in Artificial Intelligence and Security. ICAIS 2021. Communications in Computer and Information Science. Vol 1422. Springer; 2021:476-487. [doi: 10.1007/978-3-030-78615-1_42]

8. Tang G, Liu T, Cai X, Gao S, Fu L. Standardization of clinical terminology based on hybrid recall and ERNIE. Presented at: ISAIMS 2022; Oct 13-15, 2022:19-23; Amsterdam, Netherlands. URL: https://dl.acm.org/doi/10.1145/3570773.3570782 [Accessed 2026-08-28] [doi: 10.1145/3570773.3570782]
9. Zhou L, Wu CY, Wang XT, Liu SQ, Zhang YZ, Sun YM, et al. Traditional Chinese medicine synonymous term conversion: a bidirectional encoder representations from transformers-based model for converting synonymous terms in traditional Chinese medicine. World J Tradit Chin Med. 2023;9(2):224-233. [doi: 10.4103/2311-8571.378171]
10. Hu H, Cheng C, Ye Q, Peng L, Shen Y. Enhancing traditional Chinese medicine diagnostics: integrating ontological knowledge for multi-label symptom entity classification. Math Biosci Eng. Jan 2024;21(1):369-391. [doi: 10.3934/mbe.2024017] [Medline: 38303427]
11. Crammer K, Dredze M, Ganchev K, Talukdar PP, Carroll S. Automatic code assignment to medical text. In: Cohen KB, Demner-Fushman D, Friedman C, Hirschman L, Pestian J, editors. Presented at: Workshop on BioNLP 2007: biological, translational, and clinical language processing; Jun 29, 2007:129-136; Prague, Czech Republic. URL: https://aclanthology.org/W07-1017/ [Accessed 2026-08-28] [doi: 10.3115/1572392.1572416]
12. Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics. Apr 11, 2008;9(Suppl 3):S10. [doi: 10.1186/1471-2105-9-S3-S10] [Medline: 18426545]
13. Koopman B, Karimi S, Nguyen A, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. BMC Med Inform Decis Mak. Jul 15, 2015;15:53. [doi: 10.1186/s12911-015-0174-2] [Medline: 26174442]
14. Lee HC, Hsu YY, Kao HY. AuDis: an automatic CRF-enhanced disease normalization in biomedical text. Database (Oxford). 2016;2016:baw091. [doi: 10.1093/database/baw091] [Medline: 27278815]
15. Ševa J, Kittner M, Roller R, Leser U. Multi-lingual ICD-10 coding using a hybrid rule-based and supervised classification approach at CLEF eHealth 2017. In: Cappellato L, Ferro N, Goeuriot L, Mandl T, editors. Presented at: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum; Sep 11-14, 2017; Dublin, Ireland. URL: https://ceur-ws.org/Vol-1866/paper_70.pdf [Accessed 2026-08-28]
16. Zhou L, Cheng C, Ou D, Huang H. Construction of a semi-automatic ICD-10 coding system. BMC Med Inform Decis Mak. Apr 15, 2020;20(1):67. [doi: 10.1186/s12911-020-1085-4] [Medline: 32293423]
17. Perotte A, Pivovarov R, Natarajan K, Weiskopf N, Wood F, Elhadad N. Diagnosis code assignment: models and evaluation metrics. J Am Med Inform Assoc. 2014;21(2):231-237. [doi: 10.1136/amiajnl-2013-002159] [Medline: 24296907]
18. Ghiasvand O, Kate RJ. UWM: a simple baseline method for identifying attributes of disease and disorder mentions in clinical text. In: Nakov P, Zesch T, Cer D, Jurgens D, editors. Presented at: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); Jun 4-5, 2015:385-388; Denver, Colorado. URL: https://aclanthology.org/S15-2066/ [Accessed 2026-05-20] [doi: 10.18653/v1/S15-2066]
19. Huang M, Han H, Wang H, Li L, Zhang Y, Bhatti UA. A clinical decision support framework for heterogeneous data sources. IEEE J Biomed Health Inform. Nov 2018;22(6):1824-1833. [doi: 10.1109/JBHI.2018.2846626] [Medline: 29994279]
20. Xu K, Lam M, Pang J, Gao X, Band C, Mathur P, et al. Multimodal machine learning for automated ICD coding. In: Doshi-Velez F, Fackler J, Jung K, editors. Presented at: Proceedings of the 4th Machine Learning for Healthcare Conference (MLHC 2019); Aug 9-10, 2019:197-215; Ann Arbor, Michigan. URL: https://proceedings.mlr.press/v106/xu19a.html [Accessed 2026-05-20]
21. Boytcheva S. Automatic matching of ICD-10 codes to diagnoses in discharge letters. Presented at: Proceedings of the Second Workshop on Biomedical Natural Language Processing. Sep 15, 2011:Association for Computational Linguistics. 11-18; Hissar, Bulgaria. URL: https://aclanthology.org/W11-4203/ [Accessed 2026-08-28]
22. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. Nov 15, 2013;29(22):2909-2917. [doi: 10.1093/bioinformatics/btt474] [Medline: 23969135]
23. Ji Z, Wei Q, Xu H. BERT-based ranking for biomedical entity normalization. AMIA Jt Summits Transl Sci Proc. 2020;2020:269-277. [Medline: 32477646]
24. Wang Q, Ji Z, Wang J, et al. A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes. J Biomed Inform. May 2020;105:103418. [doi: 10.1016/j.jbi.2020.103418] [Medline: 32298846]
25. Xu D, Zhang Z, Bethard S. A generate-and-rank framework with semantic type regularization for biomedical concept normalization. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020:8452-8464; Online. URL: https://aclanthology.org/2020.acl-main.748/ [Accessed 2026-08-28] [doi: 10.18653/v1/2020.acl-main.748]
26. Kalyan KS, Sangeetha S. Target concept guided medical concept normalization in noisy user-generated texts. In: Agirre E, Apidianaki M, Vulić I, editors. Presented at: Proceedings of Deep Learning Inside Out (DeeLIO); Nov 19,

2020:64-73; Online. URL: https://aclanthology.org/2020.deelio-1.8/ [Accessed 2026-08-28] [doi: 10.18653/v1/2020.deelio-1.8]
27. Chen L, Varoquaux G, Suchanek FM. A lightweight neural model for biomedical entity linking. AAAI. 2021;35(14):12657-12665. [doi: 10.1609/aaai.v35i14.17499]
28. Lai T, Ji H, Zhai C. BERT might be overkill: a tiny but effective biomedical entity linker based on residual convolutional neural networks. In: Moens MF, Huang X, Specia L, Yih SWT, editors. Presented at: Findings of the Association for Computational Linguistics; Nov 7-11, 2021:1631-1639; Punta Cana, Dominican Republic. URL: https://aclanthology.org/2021.findings-emnlp.140/ [Accessed 2026-08-28] [doi: 10.18653/v1/2021.findings-emnlp.140]
29. Li L, Zhai Y, Gao J, Wang L, Hou L, Zhao J. Stacking-BERT model for Chinese medical procedure entity normalization. Math Biosci Eng. Jan 2023;20(1):1018-1036. [doi: 10.3934/mbe.2023047] [Medline: 36650800]
30. Huang CW, Tsai SC, Chen YN. PLM-ICD: automatic ICD coding with pretrained language models. In: Naumann T, Bethard S, Roberts K, Rumshisky A, editors. Presented at: Proceedings of the 4th Clinical Natural Language Processing Workshop; Jul 14, 2022:10-20; Seattle, WA. URL: https://aclanthology.org/2022.clinicalnlp-1.2/ [Accessed 2026-08-28] [doi: 10.18653/v1/2022.clinicalnlp-1.2]
31. Yan J, Wang Y, Xiang L, Zhou Y, Zong C. A knowledge-driven generative model for multi-implication Chinese medical procedure entity normalization. In: Webber B, Cohn T, He Y, Liu Y, editors. Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Nov 16-20, 2020:1490-1499; Online. URL: https://aclanthology.org/2020.emnlp-main.116/ [Accessed 2026-08-28] [doi: 10.18653/v1/2020.emnlp-main.116]
32. Liang M, Xue K, Ye Q, Ruan T. A combined recall and rank framework with online negative sampling for Chinese procedure terminology normalization. Bioinformatics. Oct 25, 2021;37(20):3610-3617. [doi: 10.1093/bioinformatics/btab381] [Medline: 34037691]
33. Tang J, Huang Z, Xu H, et al. Chinese clinical named entity recognition with segmentation synonym sentence synthesis mechanism: algorithm development and validation. JMIR Med Inform. Nov 21, 2024;12:e60334. [doi: 10.2196/60334] [Medline: 39622697]
34. Li M, Zhou H, Yang H, Zhang R. RT: a retrieving and chain-of-thought framework for few-shot medical named entity recognition. J Am Med Inform Assoc. Sep 1, 2024;31(9):1929-1938. [doi: 10.1093/jamia/ocae095] [Medline: 38708849]
35. Garda S, Weber-Genzel L, Martin R, Leser U. BELB: a biomedical entity linking benchmark. Bioinformatics. Nov 1, 2023;39(11):btad698. [doi: 10.1093/bioinformatics/btad698] [Medline: 37975879]
36. Luo G, Shi N, Wang G, Tang B. Contextual information contributes to biomedical named entity normalization. J Biomed Inform. May 2025;165:104806. [doi: 10.1016/j.jbi.2025.104806] [Medline: 40044019]
37. Pine M, Tompkins C. Evolution of the International Classification of Diseases—from hierarchical classification to linguistic nuance. JAMA Netw Open. Apr 18, 2024;7(4):e246474. [doi: 10.1001/jamanetworkopen.2024.6474] [Medline: 38635276]
38. Park HA, Yu SJ, Jung H. Strategies for adopting and implementing SNOMED CT in Korea. Healthc Inform Res. Jan 2021;27(1):3-10. [doi: 10.4258/hir.2021.27.1.3] [Medline: 33611871]
39. Alahmar A, AlMousa M, Benlamri R. Automated clinical pathway standardization using SNOMED CT-based semantic relatedness. Digit Health. 2022;8:20552076221089796. [doi: 10.1177/20552076221089796] [Medline: 35392252]
40. Williams AT, Bates-Jensen BM, Hodge F, Lee E, Levy-Storms L. Pressure injury pain over time among nursing home residents. Geriatr Nurs. 2024;59:362-371. [doi: 10.1016/j.gerinurse.2024.07.010] [Medline: 39127012]
41. Aucar N, Fagalde I, Zanella A, et al. Nocturia: its characteristics, diagnostic algorithm and treatment. Int Urol Nephrol. Jan 2023;55(1):107-114. [doi: 10.1007/s11255-022-03317-y] [Medline: 35945304]
42. Tu S, Ye H, Xin Y, et al. Early anuria in incident peritoneal dialysis patients: incidence, risk factors, and associated clinical outcomes. Kidney Med. Oct 2024;6(10):100882. [doi: 10.1016/j.xkme.2024.100882] [Medline: 39247762]
43. Rao SS, Manabe N, Karasawa Y, et al. Comparative profiles of lubiprostone, linaclotide, and elobixibat for chronic constipation: a systematic literature review with meta-analysis and number needed to treat/harm. BMC Gastroenterol. Jan 2, 2024;24(1):12. [doi: 10.1186/s12876-023-03104-8] [Medline: 38166671]
44. Dang HT, Tran DM, Phung TTB, et al. Promising clinical and immunological efficacy of Bacillus clausii spore probiotics for supportive treatment of persistent diarrhea in children. Sci Rep. Mar 18, 2024;14(1):6422. [doi: 10.1038/s41598-024-56627-9] [Medline: 38494525]
45. Han P, Li X, Zhang Z, et al. CMCN: Chinese medical concept normalization using continual learning and knowledge-enhanced. Artif Intell Med. Nov 2024;157:102965. [doi: 10.1016/j.artmed.2024.102965] [Medline: 39241561]
46. Wang A, Liu C, Yang J, Weng C. Fine-tuning large language models for rare disease concept normalization. J Am Med Inform Assoc. Sep 1, 2024;31(9):2076-2083. [doi: 10.1093/jamia/ocae133] [Medline: 38829731]
47. Nastou K, Koutrouli M, Pyysalo S, Jensen LJ. CoNECo: a corpus for named entity recognition and normalization of protein complexes. Bioinform Adv. 2024;4(1):vbae116. [doi: 10.1093/bioadv/vbae116] [Medline: 39411448]

48. Keloth VK, Hu Y, Xie Q, et al. Advancing entity recognition in biomedicine via instruction tuning of large language models. Bioinformatics. Mar 29, 2024;40(4):btae163. [doi: 10.1093/bioinformatics/btae163] [Medline: 38514400]
49. Xue M, Yu B, Liu T, Zhang Y, Meng E, Wang B. Porous lattice transformer encoder for Chinese NER. In: Scott D, Bel N, Zong C, editors. Presented at: Proceedings of the 28th International Conference on Computational Linguistics; Dec 8-13, 2020:3831-3841; Barcelona, Spain (Online). URL: https://aclanthology.org/2020.coling-main.340/ [Accessed 2026-08-28] [doi: 10.18653/v1/2020.coling-main.340]
50. Li X, Yan H, Qiu X, Huang X. FLAT: Chinese NER using flat-lattice transformer. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020:6836-6842; Online. URL: https://aclanthology.org/2020.acl-main.611/ [Accessed 2026-08-28] [doi: 10.18653/v1/2020.acl-main.611]
51. Kate RJ. Clinical term normalization using learned edit patterns and subconcept matching: system development and evaluation. JMIR Med Inform. Jan 14, 2021;9(1):e23104. [doi: 10.2196/23104] [Medline: 33443483]
52. Sung M, Jeong M, Choi Y, Kim D, Lee J, Kang J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinformatics. Oct 14, 2022;38(20):4837-4839. [doi: 10.1093/bioinformatics/btac598] [Medline: 36053172]
53. Liu S, Wang A, Xiu X, Zhong M, Wu S. Evaluating medical entity recognition in health care: entity model quantitative study. JMIR Med Inform. Oct 17, 2024;12:e59782. [doi: 10.2196/59782] [Medline: 39419501]
54. Alhassan A, Schlegel V, Aloud M, Batista-Navarro R, Nenadic G. Discontinuous named entities in clinical text: a systematic literature review. J Biomed Inform. Feb 2025;162:104783. [doi: 10.1016/j.jbi.2025.104783] [Medline: 39863246]
55. Almeida JR, Silva JF, Matos S, Oliveira JL. A two-stage workflow to extract and harmonize drug mentions from clinical notes into observational databases. J Biomed Inform. Aug 2021;120:103849. [doi: 10.1016/j.jbi.2021.103849] [Medline: 34214696]
56. García-Barragán Á, Sakor A, Vidal ME, et al. NSSC: a neuro-symbolic AI system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. Med Biol Eng Comput. Mar 2025;63(3):749-772. [doi: 10.1007/s11517-024-03227-4] [Medline: 39485651]
57. Gou X, Yao J, Lai W, et al. A framework for normalized extraction of fine-grained traditional Chinese medicine symptom entities and relations. BMC Med Inform Decis Mak. Dec 6, 2025;25(1):441. [doi: 10.1186/s12911-025-03257-4] [Medline: 41353357]
58. Wang XK, Zhang HY, Wang YN, Hou WH, Wang JQ, Peng HG. A decision support system for comments-adjusted ranking of hotels. J Oper Res Soc. Sep 2, 2025;76(9):1880-1897. [doi: 10.1080/01605682.2024.2446683]
59. Hou W, Wang J, Lin Q, Wang X, Huang L. Improving clinical foundation models with multi-modal learning and domain adaptation for chronic disease prediction. IEEE J Biomed Health Inform. 2026;30(8):7359-7372. [doi: 10.1109/JBHI.2025.3595140] [Medline: 40758490]
60. Abdulnazar A, Roller R, Schulz S, Kreuzthaler M. Unsupervised SapBERT-based bi-encoders for medical concept annotation of clinical narratives with SNOMED CT. Digit Health. 2024;10:20552076241288681. [doi: 10.1177/20552076241288681] [Medline: 39493636]
61. Liu L, Blake V, Barman M, et al. Using natural language processing to extract information from clinical text in electronic medical records for populating clinical registries: a systematic review. J Am Med Inform Assoc. Feb 1, 2026;33(2):484-499. [doi: 10.1093/jamia/ocaf176] [Medline: 41093296]

Abbreviations

AutoML: automated machine learning
BERT: Bidirectional Encoder Representations from Transformers
BiLSTM: bidirectional long short-term memory
CFFM: correlated feature fusion module
CRF: conditional random field
DRG: diagnosis-related group
DSG: directional skip-gram
EMR: electronic medical record
ERNIE: enhanced representation through knowledge integration
FLAT: flat-lattice transformer
FN: false negative
FP: false positive
HLT: hierarchical labeling tree
***ICD-10*:** *International Statistical Classification of Diseases, 10th Revision*
MIMIC-III: Medical Information Mart for Intensive Care III

MTCG: multi-task candidate generator
NEN: named entity normalization
NER: named entity recognition
NLP: natural language processing
OCMH: output comparison of multiple heterogeneous models
PLM-*ICD*: pretrained language model framework for *International Classification of Diseases* coding
PLTE: porous lattice transformer encoder
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
STC-NF: split-then-concatenate normalization framework
STC-TC: synonymous term conversion with thinking of candidate terms
symNormHS: symptom terminology normalization approach with hierarchical semantics
TCM: traditional Chinese medicine
TF-IDF: term frequency-inverse document frequency
TN: true negative
TP: true positive
WHO: World Health Organization

Edited by Arriel Benis; peer-reviewed by Jian-Qiang Wang, Tong Ruan; submitted 13.Oct.2025; final revised version received 21.Jun.2026; accepted 12.Aug.2026; published 25.Sep.2026

Please cite as:
Yao J, Gou X, Lai W, Gao Y, Wang S, Zhou C, Ye H, Tian J, Yi J, Cao D
Symptom Terminology Normalization in Traditional Chinese Medicine: Development and Evaluation of a 2-Stage Deep Learning Framework Based on Fine-Grained Semantic Classification
JMIR Med Inform 2026;14:e85825
URL: https://medinform.jmir.org/2026/1/e85825
doi: 10.2196/85825